RATAN TATA LIBRARY

DELHI UNIVERSITY LIBRARY SYSTEM

WARNING :

Please point out to the Counter Staff physical damage to the book before borrowing, if any.

RATAN TATA LIBRARY, (DULS)

(Delhi School of Economics)

Cl. No. B2805 JO

Ac. No. 7457

This book should be returned on or before the date last stamped below. An overdue charge of 10 Paise per day on general books and 25 Paise as per day on text books will be charged for the first two days and 50 Paise from third day the book is kept overrime.

12 FEB 1988

13 JAN 1992

STATISTICS IN PSYCHOLOGY AND EDUCATION

BY

HENRY E. GARRETT, Ph.D.

PROFESSOR OF PSYCHOLOGY, COLUMBIA UNIVERSITY

WITH AN INTRODUCTION BY

R. S. WOODWORTH

PROFESSOR EMERITUS OF PSYCHOLOGY

COLUMBIA UNIVERSITY

THIRD EDITION

LONGMANS, GREEN AND CO.

NEW YORK · LONDON · TORONTO

LONGMANS, GREEN AND CO., INC.
55 FIFTH AVENUE, NEW YORK 3

LONGMANS, GREEN AND CO. LTD.
6 & 7 CLIFFORD STREET, LONDON W 1

LONGMANS, GREEN AND CO.
215 VICTORIA STREET, TORONTO 1

GARRETT
STATISTICS IN PSYCHOLOGY AND EDUCATION

COPYRIGHT • 1926, 1937, AND 1947
BY LONGMANS, GREEN AND CO., INC.

ALL RIGHTS RESERVED, INCLUDING THE RIGHT TO REPRODUCE
THIS BOOK, OR ANY PORTION THEREOF, IN ANY FORM

FIRST EDITION, JANUARY 1926
TEN PRINTINGS
SECOND EDITION, REWRITTEN JUNE 1937
EIGHT PRINTINGS
THIRD EDITION, REWRITTEN JANUARY 1947
JULY 1947, NOVEMBER 1947
OCTOBER 1948, SEPTEMBER 1949
JUNE 1950

Printed in the United States of America
VAN REES PRESS • NEW YORK

INTRODUCTION

Modern problems and needs are forcing statistical methods and statistical ideas more and more to the fore. There are so many things we wish to know which cannot be discovered by a single observation, or by a single measurement. We wish to envisage the behavior of a man who, like all men, is rather a variable quantity, and must be observed repeatedly and not once for all. We wish to study the social group, composed of individuals differing one from another. We should like to be able to compare one group with another, one race with another, as well as one individual with another individual, or the individual with the norm for his age, race or class. We wish to trace the curve which pictures the growth of a child, or of a population. We wish to disentangle the interwoven factors of heredity and environment which influence the development of the individual, and to measure the similarly interwoven effects of laws, social customs and economic conditions upon public health, safety and welfare generally. Even if our statistical appetite is far from keen, we all of us should like to know enough to understand, or to withstand, the statistics that are constantly being thrown at us in print or conversation — much of it pretty bad statistics. The only cure for bad statistics is apparently more and better statistics. All in all, it certainly appears that the rudiments of sound statistical sense are coming to be an essential of a liberal education.

Now there are different orders of statisticians. There is, first in order, the mathematician who invents the method for performing a certain type of statistical job. His interest, as a mathematician, is not in the educational, social or psychological problems just alluded to, but in the problem of devising instruments for handling such matters. He is the tool-maker of the

statistical industry, and one good tool-maker can supply many skilled workers. The latter are quite another order of statisticians. Supply them with the mathematician's formulas, map out the procedure for them to follow, provide working charts, tables and calculating machines, and they will compute from your data the necessary averages, probable errors and correlation coefficients. Their interest, as computers, lies in the quick and accurate handling of the tools of the trade. But there is a statistician of yet another order, in between the other two. His primary interest is psychological, perhaps, or it may be educational. It is he who has selected the scientific or practical problem, who has organized his attack upon the problem in such fashion that the data obtained can be handled in some sound statistical way. He selects the statistical tools to be employed, and, when the computers have done their work, he scrutinizes the results for their bearing upon the scientific or practical problem with which he started. Such an one, in short, must have a discriminating knowledge of the kit of tools which the mathematician has handed him, as well as some skill in their actual use.

The reader of the present book will quickly discern that it is intended primarily for statisticians of the last-mentioned type. It lays out before him the tools of the trade; it explains very fully and carefully the manner of handling each tool; it affords practice in the use of each. While it has little to say of the tool-maker's art, it takes great pains to make clear the use and limitations of each tool. As any one can readily see who has tried to teach statistics to the class of students who most need to know the subject, this book is the product of a genuine teacher's experience, and is exceptionally well adapted to the student's use. To an unusual degree, it succeeds in meeting the student upon his own ground.

R. S. Woodworth

Columbia University
(1926)

PREFACE
To Third Edition

In this edition much of the text has been rewritten and various procedures brought up to date. Earlier chapters dealing with the frequency distribution have been changed the least, later chapters dealing with sampling and correlation have been changed the most. Several methods and formulas of limited application have been omitted in favor of more useful techniques. The new material includes small sample methods; a chapter (Chapter VIII) dealing with the testing of experimental hypotheses; a more complete treatment of the Chi-square test; an introduction to analysis of variance; and the Wherry-Doolittle method of test selection.

As before, I am indebted to Dean J. F. Walker of the University of Arizona and to Professor Vernon W. Lemmon of Washington University for advice and suggestions of various sorts. My colleagues, Dr. W. N. Schoenfeld, Dr. Joseph Zubin, and Mr. Ralph F. Hefferline, have read most of the manuscript and have offered many constructive criticisms.

HENRY E. GARRETT

COLUMBIA UNIVERSITY
(1946)

TO THE INSTRUCTOR

This book contains more material than can, perhaps, be covered thoroughly in a one semester course. The following selection of topics is suggested, therefore, as meeting the requirements of a course in "minimum essentials."

Chapters I, II, and III
Chapter IV (I and II)
Chapter V (I and II)
Chapter VI (II)
Chapter VII (I, II, III, and IV)
Chapter VIII (I and II)
Chapter IX
Chapter X (I and II)
Chapter XI (I)
Chapter XIII (I and II)

CONTENTS

CHAPTER I

THE FREQUENCY DISTRIBUTION

CHAPTER II

MEASURES OF CENTRAL TENDENCY

CHAPTER III

MEASURES OF VARIABILITY

CHAPTER IV

CUMULATIVE DISTRIBUTIONS, GRAPHIC METHODS, AND PERCENTILES

CHAPTER V

THE NORMAL PROBABILITY CURVE

CHAPTER VI

APPLICATIONS OF THE NORMAL PROBABILITY CURVE

CHAPTER VII

SAMPLING AND RELIABILITY

CHAPTER VIII

TESTING EXPERIMENTAL HYPOTHESES

CHAPTER IX

LINEAR CORRELATION

CHAPTER X

REGRESSION AND PREDICTION

CHAPTER XI

FURTHER METHODS OF CORRELATION

CHAPTER XII

THE RELIABILITY AND VALIDITY OF TEST SCORES

CHAPTER XIII

PARTIAL AND MULTIPLE CORRELATION

CHAPTER XIV

MULTIPLE CORRELATION IN TEST SELECTION

STATISTICS IN PSYCHOLOGY AND EDUCATION

CHAPTER I

THE FREQUENCY DISTRIBUTION

I. MEASURES IN GENERAL

1. What Is Meant by Measurement

THE measurement of individuals and objects may be of various kinds, and may be taken to varying degrees of precision. When individuals have been *ranked* or arranged in a series with respect to some attribute or trait, we have perhaps the simplest sort of measurement. Children may be put in order for height, weight, or regularity of school attendance; salesmen may be ranked for years of experience, or amount of sales over a year; advertisements or pictures may be ranked for amount of color, or for cost, or for sales appeal. Rank order tells us, in a rough way, how much of an attribute a given person or thing possesses. But it tells us little else except serial position in a group. We cannot add or subtract ranks as we can inches or pounds: a person's rank is always relative to the ranks of other members of his group, and is never absolute, i.e., in terms of some known unit.

Measurements of individuals may also be expressed as *scores*. Scores are usually given in terms of *time* taken to complete a task, or *amount* done in a given time; less often scores are expressed in terms of difficulty of the task performed, or excellence of the final result. Scores vary with performance, although score-changes probably do not parallel performance-changes exactly. When scores are expressed in equal units, they constitute a *scale*. Scaled tests in psychology and education have equal units or steps but do *not* possess an absolute zero point. On the other hand, the "c.g.s. scales" (centimeters, grams, seconds) of physics do have equal units and an absolute

zero point. "Scores" from physical scales are called *measures;* they may be added or subtracted and a "score" of twenty inches, say, is twice a "score" of ten inches. Scaled scores from mental tests may also be added or subtracted just as we add and subtract inches. But we cannot say that a score of 40 achieved on a test is twice as good as a score of 20, since neither is measured from a zero point of just no ability. Traits and other characteristics, determinations of which are expressible as scores or measures, are known generally as *variables.*

2. Continuous and Discrete Series

In the measurement of mental and social traits, most of the variables with which we deal fall into *continuous series.* A continuous series is one which is capable of any degree of subdivision, although in practice divisions smaller than some convenient unit are rarely employed. Measurements of general intelligence illustrate scores which fall into continuous series. I.Q.'s, for example, may be thought of as increasing by increments of 1 on an ability continuum which extends from the idiot to the genius. But there is no reason why with more refined methods of measurement we should not be able to get I.Q.'s of 100.8 or even of 100.83. Physical measures such as height, weight, and cephalic index as well as scores from mental and educational tests fall into continuous series: within the given range any measure, integral or fractional, may exist and have meaning. When gaps occur in a truly continuous series, these are to be attributed to a failure to measure enough cases, to the relative crudity of the measuring instrument, or to some other factor of a like sort, rather than to the lack of measures within the gaps.

Not all variables fall into continuous series. A salary scale in a department store may run from $10 per week to $20 per week in units of $1; no one receives, let us say, $17.53 per week. Again, the average family in a certain locality may work out mathematically to have 2.57 children, although there is obviously a real gap between two children and three children.

Series which exhibit real gaps are called *discrete* or *discontinuous.*

It is perhaps fortunate that nearly all of the variables with which we deal in psychology and education fall into continuous series or may be profitably treated as continuous. This makes it possible for us to concern ourselves for the present with methods of handling continuous data, and to postpone the discussion of discrete data to a later page (68).

In the following sections we shall define more precisely just what is meant by a score in a continuous series, and then show how scores may be classified into what is called a *frequency distribution.*

3. The Meaning of Scores in Continuous Series

Scores or other numbers in continuous series are to be thought of as *distances* along a continuum, rather than as discrete points. An inch is the linear magnitude between two divisions on a foot-rule; and, in like manner, a score in a mental test is a unit distance between two limits. A score of 150 upon an intelligence examination, for example, represents the interval 149.5 up to 150.5. The exact midpoint of this score-interval is 150 as shown below.

Score 150

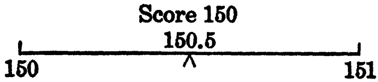
150

149.5 ^ 150.5

Other scores are to be interpreted in the same way. A score of 8 on the Thorndike Handwriting Scale, for instance, includes all values from 7.5 up to 8.5; i.e., any value from a point .5 unit *below* 8, to .5 unit *above* 8. Hence, 7.7, 8.0, and 8.4 may all be scored 8. An interval extending from .5 unit below to .5 unit above the given value is the usual mathematical meaning of a single score.

There is another and somewhat different meaning which a test score may have. According to this second view, a score of 150 means that an individual has done *at least* 150 items correctly, but not 151. Hence, a score of 150 represents any

value *between* 150 and 151. Any fractional value greater than 150, but less than 151, e.g., 150.3 or 150.8, since it falls within the interval 150–151 is scored simply as 150. The middle of the score interval is 150.5. (See below.)



Both of these ways of defining a score are valid and useful. Which to use will depend upon the way in which the test is scored and on the meaning of the units of measurement employed. If each of ten boys is recorded as having a height of sixty-four inches this will ordinarily mean that these heights fall between 63.5 and 64.5 inches (middle value 64 in.), and not between sixty-four and sixty-five inches (middle value 64.5 in.). On the other hand, the ages of twenty-five children, all recorded as being nine years old, will most probably lie between nine and ten years; will be greater than nine and less than ten years (middle value 9.5). But "nine years old" must be taken in many studies to mean 8.5 up to 9.5 years with a middle value of nine years. The point to remember is that results obtained from treating scores under our second definition will always be .5 unit higher than results obtained when scores are taken under the first or mathematical definition. The student will often have to decide, perhaps somewhat arbitrarily, which meaning a score should have. As a general rule it is safer to take the first meaning of a score unless clearly indicated otherwise. This will be the method followed throughout this book. That is, scores of 62 and 231, say, will usually mean 61.5 up to 62.5, and 230.5 up to 231.5, and not 62 up to 63, and 231 up to 232.

II. Drawing Up a Frequency Distribution

1. The Classification of Measures

Data collected from tests and experiments often have little meaning or significance until they have been rearranged or

classified in a systematic way. The first task that confronts us, then, is the organization of our material and this leads naturally to a grouping of the measures or scores into classes or categories. The procedure in grouping falls under three main heads:

(1) Determination of the *range* or the interval between the largest and smallest scores. The range is found by subtracting the smallest from the largest score.

(2) Decision as to the *number* and *size* of the groups to be used in classification. The number and size of these *class-intervals* will depend upon the range of scores and the kind of measures with which we are dealing.

(3) Tabulation of the separate scores within their proper class-intervals.

These three principles of classification are illustrated in Table 1. The figures in this table represent the Army Alpha scores earned by fifty college men. Since the highest score is 197, and the lowest 142, the range (197–142) is exactly 55. In deciding upon the number of classes to be used in grouping, a good general rule is to select by trial an interval which will yield not more than twenty nor less than ten classes.*

The number of class-intervals which a given range will yield can be determined approximately (within one interval) by dividing the range by the interval tentatively chosen. In the present problem, 55 (the range) divided by 5 (the interval) gives 11, which is one less than the actual number of intervals, namely, 12. An interval of three units will yield nineteen classes; an interval of ten units, six classes.

The tabulation of the separate scores within their class-intervals is shown in Table 1. In the first column of this table the class-intervals have been listed serially from the smallest score at the bottom of the column to the largest score at the top. Each class-interval comprises exactly five scores. The first interval "140 up to 145" begins with score 140 and ends with 144, thus including the five scores 140, 141, 142, 143, and

* This rule must often be broken when the number of scores is very large or very small.

TABLE 1

THE TABULATION OF ARMY ALPHA SCORES MADE BY FIFTY COLLEGE STUDENTS

1. The original scores ungrouped

185	166	176	145	166	191	177	164	171	174
147	178	176	# 142	170	158	171	167	180	178
173	148	168	187	181	172	165	169	173	184
175	156	158	187	156	172	162	193	173	183
* 197	181	151	161	153	172	162	179	188	179

* Highest score # Lowest score

2. The same fifty scores grouped into a frequency distribution

(1) Class-Intervals	(2) Tallies	(3) f (frequency)
195 up to 200	/	1
190 " " 195	//	2
185 " " 190	////	4
180 " " 185	~~////~~	5
175 " " 180	~~////~~ ///	8
170 " " 175	~~////~~ ~~////~~	10
165 " " 170	~~////~~ /	6
160 " " 165	////	4
155 " " 160	////	4
150 " " 155	//	2
145 " " 150	///	3
140 " " 145	/	1
		$N = 50$

144. The second interval "145 up to 150" begins with 145 and ends with 149, i.e., at score 150. The last interval "195 up to 200" begins with score 195 and ends at score 200, thus including the scores 195, 196, 197, 198, 199. In column (2), marked "Tallies," the separate scores have been listed opposite their proper intervals. The first score, 185, is represented by a tally placed opposite interval "185 up to 190"; the second score, 147, by a tally placed opposite interval "145 up to 150"; and the third score, 173, by a tally placed opposite "170 up to 175." The remaining scores have been tabulated in the same way. When all fifty scores have been listed, the total number of tallies on each class-interval (i.e., the frequency) is written in column (3) headed f (frequency). The sum of the f column is

called N. When the total frequency within each class-interval has been tabulated opposite the proper interval, as shown in column (3), our fifty Army Alpha scores are arranged in a *frequency distribution.*

The reader will note that the beginning score of the first interval in the distribution (140 up to 145) has been set at 140 although the lowest score in the series is 142. When the interval selected for tabulation is five units it facilitates tabulation as well as computations which come later if the score limits of the first interval, and, accordingly, of each successive interval, are multiples of five. A class-interval "142 up to 147" is just as good theoretically as a class-interval "140 up to 145"; but the second is easier to handle from the standpoint of the arithmetic involved.

2. Methods of Describing the Limits of the Class-Intervals in a Frequency Distribution

Table 2 illustrates three ways of expressing the limits of the class-intervals in a frequency distribution. In (A), the interval "140 up to 145" means, as we have already seen, that all scores from 140 up to but not including 145 fall within this grouping. The intervals in (B) cover the same distances as in (A), but the upper and lower limits of each interval are defined more exactly. We have seen (p. 3) that a score of 140 in a continuous series ordinarily means the interval 139.5 up to 140.5; and that a score of 144 means 143.5 up to 144.5. Accordingly, to express precisely the fact that an interval *begins* with 140 and *ends* with 144, we may write 139.5 (the beginning of score 140) as the lower limit, and 144.5 (end of score 144 or beginning of score 145) as the upper limit of this step. The class-intervals in (C) express the same facts more clearly than in (A) and less exactly than in (B). Thus, "140–144" means that this interval begins *with* score 140 and ends *with* score 144; but the precise limits of the interval are not given. The diagram below will show how (A), (B), and (C) are three ways of expressing identically the same facts:

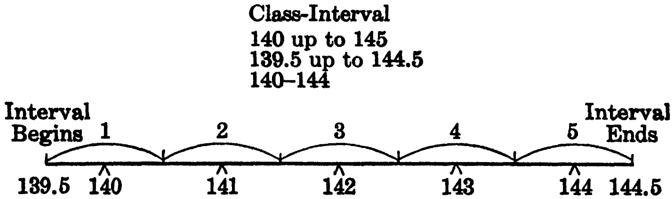


TABLE 2

METHODS OF GROUPING SCORES INTO A FREQUENCY DISTRIBUTION

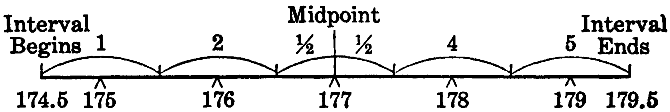
(The data are the fifty Army Alpha scores tabulated in Table 1, p. 6)

(*A*) Class-Intervals	Mid-point	*f*	(*B*) Class-Intervals	Mid-point	*f*	(*C*) Class-Intervals	Mid-point	*f*
195 up to 200	197	1	194.5 up to 199.5	197	1	195–199	197	1
190 " " 195	192	2	189.5 " " 194.5	192	2	190–194	192	2
185 " " 190	187	4	184.5 " " 189.5	187	4	185–189	187	4
180 " " 185	182	5	179.5 " " 184.5	182	5	180–184	182	5
175 " " 180	177	8	174.5 " " 179.5	177	8	175–179	177	8
170 " " 175	172	10	169.5 " " 174.5	172	10	170–174	172	10
165 " " 170	167	6	164.5 " " 169.5	167	6	165–169	167	6
160 " " 165	162	4	159.5 " " 164.5	162	4	160–164	162	4
155 " " 160	157	4	154.5 " " 159.5	157	4	155–159	157	4
150 " " 155	152	2	149.5 " " 154.5	152	2	150–154	152	2
145 " " 150	147	3	144.5 " " 149.5	147	3	145–149	147	3
140 " " 145	142	1	139.5 " " 144.5	142	1	140–144	142	1
	$N =$	50		$N =$	50		$N =$	50

For the rapid tabulation of scores within their proper intervals, method (*C*) is to be preferred to (*B*) or (*A*). In (*A*) it is fairly easy, even when one is on guard, to let a score of 160, say, slip into the interval "155 up to 160," owing simply to the presence of 160 at the upper limit of the interval. Method (*B*) is clumsy and time-consuming because of the need for writing .5 at the beginning and end of every interval. Method (*C*), while easiest for tabulation, offers the difficulty that in later calculations one must constantly remember that the *expressed* class limits are not the *actual* class limits: that interval "140–144" begins at 139.5 (not 140) and ends at 144.5 (not 144). If this is clearly understood, method (*C*) is as accurate as (*B*) or (*A*). It will be generally used throughout this book.

The scores grouped within a given interval in a frequency distribution are assumed to be spread evenly over the entire

interval. This assumption is made whether the interval is three, five, or ten units. If we wish to represent *all* of the scores within a given interval by some single value, the midpoint of the interval is taken to be the logical choice. For example, in the interval 175–179 [Table 2, method (*C*)] all of the eight scores upon this interval are represented by the single value 177, the midpoint of the interval.* Why 177 is the midpoint of this interval is shown graphically below:



A simple rule for finding the midpoint of an interval is

$$\text{Midpoint} = \text{lower limit of interval} + \frac{(\text{upper limit} - \text{lower limit})}{2}.$$

In our illustration, $174.5 + \frac{(179.5 - 174.5)}{2} = 177$. Since the interval is five units, it follows that the midpoint must be 2.5 units from the *lower limit* of the class, i.e., $174.5 + 2.5$; or 2.5 units from the *upper limit* of the class, i.e., $179.5 - 2.5$.

It is often a question whether the midpoint is, in fact, fairly representative of *all* of the scores upon a given interval. Referring to Table 1, we find that of the ten scores in the class-interval "170 up to 175" (midpoint 172), three (170, 171, 171) are *below* the midpoint; three (172, 172, 172) are *on* the midpoint; and four (173, 173, 173, 174) are *above* the midpoint. Of the five scores upon interval "180 up to 185," three (180, 181, 181) are *below* the midpoint (182); and two (183, 184) are *above*. The single score of 197 upon interval "195 up to 200" falls exactly on the midpoint. In these examples the midpoint represents quite adequately the scores within the given intervals; but it must be admitted that the balancing of scores above and below the midpoint is not always so satisfactory as it is here. When the data are scanty, or when the distribution is badly

* The same value (namely, 177) is, of course, the midpoint of the interval when methods (*A*) and (*B*) are used.

skewed (p. 119), there may be many more scores on one side of a midpoint than on the other. When this happens, the midpoint does not fairly represent *all* of the scores within the given interval.

The assumption that the midpoint is the most representative score within an interval holds best when the number of scores in the distribution is large, and when the intervals are not too broad. But even when neither of these conditions fully obtains, the midpoint assumption is not greatly in error and is the best that we can make. In the long run, about as many scores will fall above as below the various midpoint values; and lack of balance in one interval will usually be offset by the opposite condition in another interval.

Measures of central tendency (p. 32) and of variability (p. 49) calculated from data grouped into intervals of five units, say, will usually vary slightly from the same measures calculated from these data when ungrouped, or when grouped into intervals of, say, three or ten units. These variations arise from (1) differences in the size of the groups in which the data are classified, and (2) the fact that each score within an interval is assigned the value of the middle of the interval instead of its actual value. Corrections are sometimes applied to the measures of variability to correct the *grouping error* thus introduced. But usually the error which results from grouping is so small that it may be neglected in ordinary statistical work.

III. The Graphic Representation of the Frequency Distribution

Aid in analyzing numerical data may often be obtained from a graphic or pictorial treatment of the frequency distribution. The advertiser has long used graphic methods because these devices catch the eye and hold the attention when the most careful array of statistical evidence fails to attract notice. For this and other reasons the research worker also utilizes the attention-getting power of visual presentation; and, at the same time, seeks to translate numerical facts — often abstract

and difficult of interpretation — into more concrete and understandable form.

Four methods of representing a frequency distribution graphically are in general use. These methods yield the *frequency polygon*, the *histogram*, the *cumulative frequency graph*, and the *cumulative percentage curve* or *ogive*. The first two graphic devices will be treated in the following sections; the second two in Chapter V.

1. Graphical Representation of Data; General Principles

Before considering methods of constructing a frequency polygon or histogram, we shall review briefly the simple algebraic principles which apply to all graphical representation of data. Graphing or plotting is done with reference to two lines or *coördinate axes*, the one the vertical or *Y-axis*, the other the horizontal or *X-axis*. These basic lines are perpendicular to each other, the point where they intersect being called O, or the *origin*. Figure 1 represents a system of coördinate axes.

The origin is the zero point or point of reference for both axes. Distances measured along the *X-axis* to the *right* of O are called positive, distances measured along the *X-axis* to the *left* of O negative. In the same way, distances measured on the *Y-axis above* O are positive; distances *below* O negative. By their intersection at O, the *X*- and *Y-axes* form four divisions or quadrants. In the upper right division or first quadrant (see Fig. 1), both x and y measures are positive (+ +). In the upper left division or second quadrant, x is minus and y plus (− +). In the lower left or third quadrant, both x and y are negative (− −); while in the lower right or fourth quadrant, x is plus and y minus (+ −).

To locate or plot a point "A" whose coördinates are $x = 4$, and $y = 3$, we go out from O four units on the *X-axis*, and up from the origin three units on the *Y-axis*. Where the perpendiculars to these points intersect, we locate the point "A" (see Fig. 1). The point "B," whose coördinates are $x = -5$, and $y = -7$, is plotted in the third quadrant by going left from O

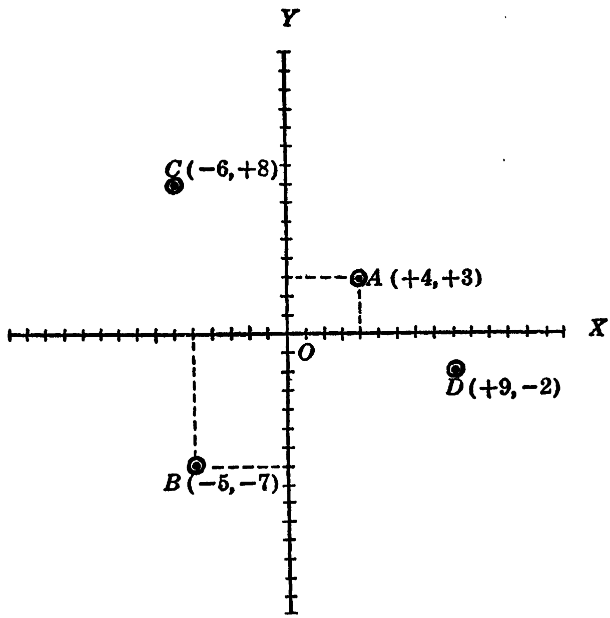


FIG. 1. A System of Coördinate Axes.

along the *X-axis* five units, and then down seven units, as shown in the figure. In like manner, any points "*C*" and "*D*" whose x and y values are known can be located with reference to *OY* and *OX*, the coördinate axes. The distance of a point from *O* on the *X-axis* is commonly called the *abscissa;* and the distance of the point from *O* on the *Y-axis* the *ordinate.* The abscissa of point "*D*" is + 9, and the ordinate, − 2.

2. The Frequency Polygon

(1) Construction of the Frequency Polygon

Figure 2 illustrates the use of the coördinate system in the construction of a frequency polygon. This graph pictures the frequency distribution of the fifty Army Alpha scores shown in Table 1, page 6. The exact limits of the intervals are laid off at regular distances along the base line (the *X-axis*) from

the origin; and the frequencies within each interval are measured off upon the *Y-axis*. There is one score on the first interval, 140 up to 145 (Table 1, p. 6). To represent this score on the diagram, we go out on the *X-axis* to 142, midway between 139.5 and 144.5, and count up one Y-unit. The frequency on the next interval, 145 up to 150, is three, hence the second point falls midway between 144.5 and 149.5, three units above the *X-axis*. The two scores on interval 150 up to 155, the four scores on 155 up to 160, and the frequency on each succeeding interval, are represented in every case by a point the specified number of scores (Y-units) above the *X-axis*, and midway between the upper and lower limits of the interval upon which the f lies. It is important in plotting a frequency polygon to remember that the midpoint of an interval is always taken to represent the entire interval. The height of the ordinate at the midpoint represents *all* of the scores within the given interval.

When all of the points have been located, they are joined in regular order to give the frequency polygon * shown in Figure 2. In order to complete the figure, one interval (134.5 to 139.5) at the low end, and one interval (199.5 to 204.5) at the high end of the distribution have been included on the X-scale. The frequency on each of these intervals is zero at the midpoint; hence by including them we begin the frequency polygon one-half interval *below* the first, and end it one-half interval *above* the last, class-interval on the *X-axis*.

In order to give symmetry and balance to a polygon, one must exercise care in the selection of unit-distances to represent the intervals on the *X-axis* and the frequencies on the *Y-axis*. A too-long X-unit tends to stretch out the polygon, while a too-short X-unit crowds the separate points. On the other hand, a too-long Y-unit exaggerates the changes from interval to interval, and a too-short Y-unit makes the polygon too flat. A good general rule is to select X- and Y-units which will make the *height* of the figure approximately 75% of its *width*. The

* Polygon means "many-sided figure."

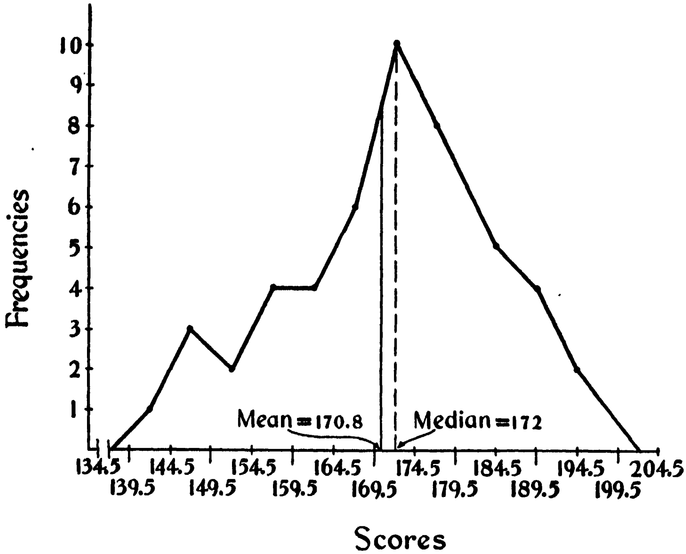


FIG. 2. Frequency Polygon Plotted from the Distribution of Fifty Army Alpha Scores Given in Table 1, page 6.

TABLE 3

SCORES MADE BY 200 ADULTS UPON A CANCELLATION TEST
Class-Interval = 4

Class-Intervals Scores	Midpoint X	f
135.5 up to 139.5	137.5	3
131.5 " " 135.5	133.5	5
127.5 " " 131.5	129.5	16
123.5 " " 127.5	125.5	23
119.5 " " 123.5	121.5	52
115.5 " " 119.5	117.5	49
111.5 " " 115.5	113.5	27
107.5 " " 111.5	109.5	18
103.5 " " 107.5	105.5	7
		$N = 200$

ratio of height to width may vary from 60–80% and the figure still have good proportions; but it can rarely go below 50% and leave the figure well balanced. The frequency polygon in

Figure 2 illustrates the "75% rule." There are thirteen class-intervals laid off on the *X-axis* — twelve full intervals plus one-half interval at the beginning and at the end of the range. Hence, our polygon should be 75% of thirteen, or about ten *X-axis* units high. These ten units (each equal to *one* interval) are laid off on the *Y-axis*. To determine how many scores (f's) should be assigned to *each unit* on the *Y-axis*, we divide 10, the largest f (on interval 169.5 up to 174.5) by 10, the number of intervals laid off on Y. The result (i.e., 1) shows that each Y-unit is exactly equal to one f or score, as shown in Figure 2.

The polygon in Figure 5, page 20, furnishes another illustration of this method of plotting a frequency polygon so as to preserve balance. This polygon represents the distribution of 200 cancellation scores shown in Table 3. There are ten intervals laid off along the base line or *X-axis* — nine full intervals plus one-half interval at the beginning and at the end of the range. Since 75% of 10 is 7.5, the height of our figure could be either seven or eight *X-axis* units. To determine the "best" value for each Y-unit, we divide 52, the largest f (on 119.5 up to 123.5) by 7, getting $7\frac{3}{7}$; and then by 8, getting 6.5. Using whole numbers for convenience, evidently we may lay off on the *Y-axis* seven units, each representing eight scores; or eight units each representing seven scores. The first combination was chosen because a unit of eight f's is somewhat easier to handle than one of seven. A slightly longer Y-unit representing ten f's would perhaps have been still more convenient.

The total frequency (N) of a distribution is represented by the *area* of its polygon; that is, the area bounded by the frequency surface and the *X-axis*. The area lying above any given interval, however, cannot be taken as proportional to the number of cases within the interval because of the irregularities in the distribution and consequently in the frequency surface. To show the positions of the mean and the median in the graph, we may locate these measures on the *X-axis* as shown in Figures 2 and 5. Perpendiculars erected at these

points show the approximate frequency at the mean and at the median.

Steps involved in constructing a frequency polygon may be summarized as follows:

(1) Draw two straight lines perpendicular to each other, the vertical line near the left side of the paper, the horizontal line near the bottom. Label the vertical line (the *Y-axis*) OY, and the horizontal line (the *X-axis*) OX. Put the O where the two lines intersect. This point is the *origin.*

(2) Lay off the intervals of the frequency distribution at regular distances along the *X-axis*. Begin with the lower limit of the interval *next below* the lowest in the distribution, and end with the upper limit of the interval *next above* the highest in the distribution. Label the successive X distances with the interval limits. Select an X-unit which will allow all of the intervals to be represented easily on the graph paper.

(3) Mark off on the *Y-axis* successive units to represent the scores (the frequencies) on the different intervals. Choose a Y-scale which will make the largest *frequency* (the height) of the polygon approximately 75% of the width of the figure.

(4) At the midpoint of each interval on the *X-axis* go up in the Y direction a distance equal to the number of scores on the interval. Place points at these locations.

(5) Join the points plotted in (4) with straight lines to give the frequency surface.

(2) Smoothing the Frequency Polygon

Because the sample is small ($N = 50$) and the frequency distribution somewhat irregular, the polygon in Figure 2 tends to be jagged in outline. To iron out chance irregularities, and also get a better notion of how the figure might look if the data were more numerous, the frequency polygon may be "smoothed" as shown in Figure 3, page 17. In smoothing, a series of "moving" or "running" averages are taken from which new or adjusted frequencies are determined. The method is illustrated in Figure 3. To find an adjusted or "smoothed" f, we add together the f on the given interval and the f's on the two

adjacent intervals (the one just below and the one just above) and divide the sum by 3. For example, the smoothed f for interval 174.5 up to 179.5 is $\frac{5+8+10}{3}$ or 7.67; for interval 154.5 up to 159.5, $\frac{4+4+2}{3}$ or 3.33. The smoothed f's for the other intervals may be found in the table below Figure 3. To find the smoothed f's for the two intervals at the extremes of the original distribution, namely, 139.5 up to 144.5, and

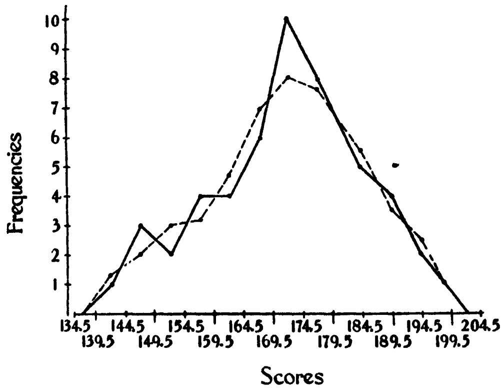


FIG. 3. Original and Smoothed Frequency Polygon. (Data from Table 1, p. 6.) The original and smoothed f's are given below.

Scores	f	Smoothed f
200–204	0	.33
195–199	1	1.00
190–194	2	2.33
185–189	4	3.67
180–184	5	5.67
175–179	8	7.67
170–174	10	8.00
165–169	6	6.67
160–164	4	4.67
155–159	4	3.33
150–154	2	3.00
145–149	3	2.00
140–144	1	1.33
135–139	0	.33
	50	50.00

194.5 up to 199.5, a slightly different procedure is necessary. Here we add 0, the f on the step *below* or *above*, the f on the given step, and the f on the adjacent step and divide by 3. This procedure makes the smoothed f for 139.5 up to 144.5, $\frac{0+1+3}{3}$ or 1.33, and the smoothed f for 194.5 up to 199.5, $\frac{2+1+0}{3}$ or 1.00. The smoothed f for the intervals 134.5 up to 139.5 and 199.5 up to 204.5, for which the frequency in the original distribution is 0, is in each case $\frac{1+0+0}{3}$ or .33. Note that if we omit these two intervals the N for the smoothed distribution will be less than 50, since the smoothed distribution has frequencies outside the range of the original distribution.

If the already smoothed f's in Figure 3 are subjected to a second smoothing, the outline of the frequency surface will become more nearly a continuous flowing curve. It is doubtful, however, whether so much adjustment of the original f's is often warranted. When an investigator presents only the smoothed frequency polygon and does not give his original data, it is impossible for a reader to tell with what he started. Moreover, smoothing gives a picture of what an investigator *might* have gotten (not what he did get) if his data had been more numerous, or less subject to error than they were. If N is large, smoothing may not greatly change the shape of a graph, and hence is often unnecessary. The frequency polygon in Figure 5, page 20, for example, which represents the distribution of 200 cancellation test scores, is quite regular without any adjustment of the ordinate (i.e., the Y) values. Probably the best course for the beginner to follow is to smooth data as little as possible. When smoothing seems to be indicated in order better to bring out the facts, one should be careful always to present original data along with "adjusted" results.

8. The Histogram or Column Diagram

A second way of representing a frequency distribution graphically is by means of a histogram or column diagram. This type of graph is illustrated in Figure 4, page 20, for the same distribution of scores represented by the frequency polygon in Figure 3, page 17. The two figures are constructed in much the same way, with this important difference: In a frequency polygon all of the scores within a given interval are represented by the midpoint of that interval, while in a histogram the assumption is made that scores are spread uniformly over their intervals. The measures within each interval of a histogram, therefore, are represented by a rectangle, the base of which equals the interval, and the height of which equals the number of scores (the f) within the interval. Thus the one score upon interval 139.5 up to 144.5 is represented by a rectangle whose base equals the length of the interval, and whose height equals one unit measured off on the Y-*axis*. The three scores within the next interval, 144.5 up to 149.5, are represented by a rectangle one interval long and three Y-units high. The altitudes of the other rectangles vary with the number of f's upon the intervals, the bases all being one interval long. When the same number of scores falls within two or more adjacent intervals, as in the intervals 154.5 up to 159.5, and 159.5 up to 164.5, the top of the rectangle covers two or more intervals on the X-*axis*. The highest rectangle is, of course, that one (on interval 169.5 up to 174.5) which has 10, the largest frequency, as its altitude. In selecting scales for the X- and Y-*axes*, the same considerations, as to height and width of figure, outlined on page 13 for the frequency polygon, should be observed.

Although in a histogram each interval is represented by a separate rectangle, it is not necessary to project the sides of the rectangles to the base line as is done in Figure 4, page 20. The rise or fall of the boundary line shows the increase or decrease in the number of scores from interval to interval and is usually the important fact to be brought out (see Fig. 5). As in a frequency polygon, the total frequency (N) is represented

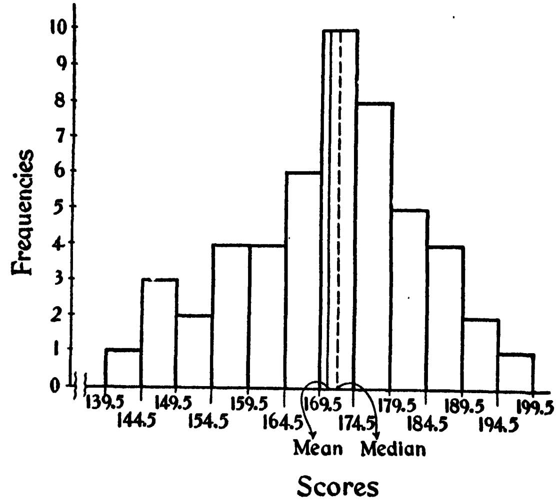


FIG. 4. Histogram of the Fifty Army Alpha Scores Shown in Table 1, page 6.

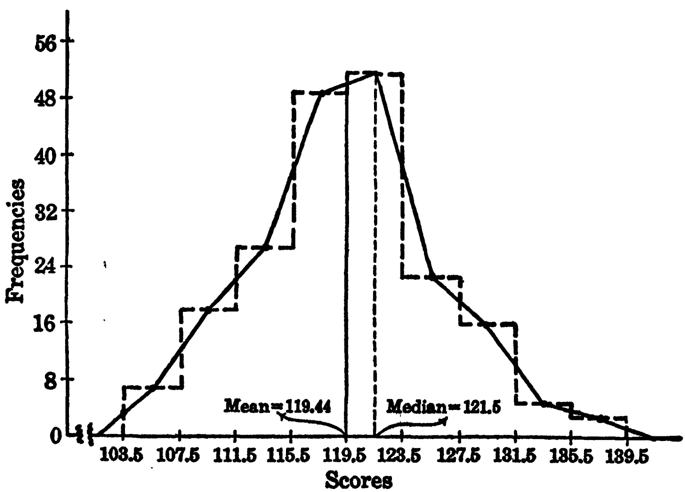


FIG. 5. Frequency Polygon and Histogram of 200 Cancellation Scores Shown in Table 3, page 14.

by the *area* of the histogram. In contrast to the frequency polygon, however, the area of *each rectangle* in a histogram is directly proportional to the number of measures within the interval. For this reason, the histogram presents an accurate picture of the relative proportions of the total frequency from interval to interval.

In order to provide a more detailed comparison of the two types of frequency graph, the distribution in Table 3, page 14, is plotted upon the same coördinate axes in Figure 5, page 20, as a frequency polygon and as a histogram. The increased number of cases and the more symmetrical arrangement of scores in the distribution make these figures more regular in appearance than those in Figures 2 and 4, pages 14 and 20.

4. Plotting Two Frequency Distributions on the Same Axes, When Samples Differ in Size

Table 4 gives the distributions of scores on an achievement examination made by two groups, A and B, which differ considerably in size. Group A has 60 cases, Group B, 160 cases.

TABLE 4

(1) Achievement Examination Scores	(2) Group A f	(3) Group B f	(4) Group A Percent-Frequencies	(5) Group B Percent-Frequencies
80–89	0	9	0.0	5.6
70–79	3	12	5.0	7.5
60–69	10	32	16.7	20.0
50–59	16	48	26.7	30.0
40–49	12	27	20.0	17.0
30–39	9	20	15.0	12.5
20–29	6	12	10.0	7.5
10–19	4	0	6.7	0.0
	60	160	100.1	100.1

If the two distributions in Table 4 are plotted as polygons or as histograms on the same coördinate axes, the fact that the f's of Group B are so much larger than those of Group A makes it hard to compare directly the range and quality of achievement

in the two groups. A useful device in cases where the N's differ in size is to express both distributions in percentage frequencies as shown in Table 4. Both N's are now 100, and the f's are comparable from interval to interval. For example, we know at once that 26.7% of Group A and 30% of Group B made scores of 50 through 59, and that 5% of the A's and 7.5% of the B's scored from 70 to 79. Frequency polygons repre-

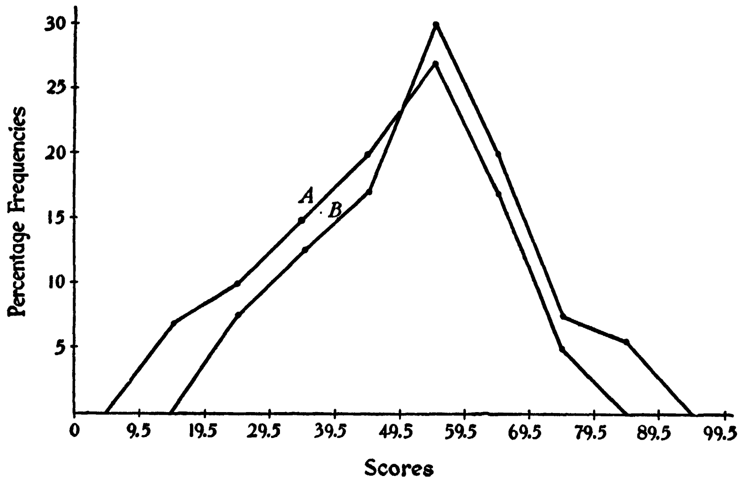


FIG. 6. Frequency Polygons of the Two Distributions in Table 4. Scores are laid off on the *X-axis*, percentage frequencies on the *Y-axis*.

senting the two distributions, in which percentage frequencies instead of original f's have been plotted on the same axes, are shown in Figure 6. These polygons provide an immediate comparison of the relative achievement of our two groups not given by polygons plotted from original frequencies.

Percentage frequencies are readily found by dividing each f by N and multiplying by 100. Thus $3/60 \times 100 = 5.0$. A simple method of finding percentage frequencies when a calculating machine is available is to divide 100 by N and, putting this figure in the machine, to multiply each f in turn by it. For

example: 1.667 (i.e., 100/60) $\times$ 3 = 5.0; 1.667 $\times$ 10 = 16.7, etc.; .625 (i.e., 100/160) $\times$ 9 = 5.6, .625 $\times$ 12 = 7.5, etc. What percentage frequencies do, in effect, is to scale each distribution down to the same total N of 100, thus permitting a comparison of f's for each interval.

5. When to Use the Frequency Polygon and When to Use the Histogram

The question of when to use the frequency polygon and when to use the histogram cannot be answered by a general rule which will cover all cases. The frequency polygon is less exact than the histogram in that it does not represent accurately, i.e., in terms of area, the number of measures within successive intervals. In comparing two or more graphs plotted on the same axes, however, the frequency polygon is the more useful, since the vertical and horizontal lines in the two histograms will often coincide. Both the histogram and the frequency polygon tell the same story and both are useful in enabling us to show in graphic form whether the scores of a group are distributed symmetrically or whether they are piled up at the low or at the high end of the scale. Not only information with regard to the group, but information with regard to the test, may be secured from a graph. If a test is too easy, the scores will crowd the high end of the scale; if the test is too hard, the scores will pile up at the low end of the scale. If the test is well suited to the group, scores will tend to be distributed symmetrically around the mean, a few individuals scoring high, a few low, and the majority scoring somewhere near the middle of the scale. When this happens, the frequency graph approximates the "ideal" or normal frequency curve described in Chapter V.

IV. Standards of Accuracy in Computation *

"How many places" to carry numerical results is a question which arises persistently in statistical computation. Sometimes

* This section should be reviewed frequently, and referred to in solving the problems given in succeeding chapters.

a student, by discarding decimals, throws away legitimate data. More often, however, he tends to retain too many decimals, a practice which may give a false appearance of great precision not always justified by the original material.

In this section are given some of the generally accepted principles which apply to statistical calculation. Observance of these rules will lead to greater uniformity in calculation. They should be followed carefully in solving the problems given in this book.

1. Rounded Numbers

In calculation, numbers are usually "rounded" off to the standard of accuracy demanded by the problem. If we round off 8.6354 to two decimals it becomes 8.64; to one decimal, 8.6; to the nearest integer, 9. Measures of central tendency and variability, coefficients of correlation, and other measures, are rarely reported to more than two decimal places. A mean of 52.6872, for example, is usually reported as 52.69; a standard deviation of 12.3841 as 12.38; and a coefficient of correlation of .6350 as .63, etc. It is very doubtful whether much of the work in mental measurement warrants accuracy beyond the second decimal. Convenient rules for rounding numbers to two decimals are as follows: When the third decimal is less than 5, drop it; when greater than 5, increase the preceding figure by 1; when exactly 5, compute the fourth decimal and correct back to the second place; when exactly 5 followed by zeros. drop it and make no correction.

2. Significant Figures

The measurement 64.3 inches is assumed to be correct to the nearest tenth of an inch, its true value lying somewhere between 64.25 and 64.35 inches. Two places to the left of the decimal point, and one to the right are fixed, and hence 64.3 is said to contain *three* significant figures. The numbers 643 and .643 also contain three significant figures each.

In the number .003046 there are *four* significant figures,

3, 0, 4, and 6, the first two zeros serving merely to locate the decimal point. When used to locate a decimal point only, a zero is not considered to be a significant figure; .004, for example, has only *one* significant figure, the two zeros simply fixing the position of 4, the significant digit. The following illustrations should make clear the matter of significant figures:

136 has *three* significant figures.

136,000 has *three* significant figures also. The true value of this number lies between 136,500 and 135,500. Only the first three digits are definitely fixed, the zeros serving simply to locate the decimal point or fix the size of the number.

1360. has *four* significant figures; the decimal indicates that the zero in the fourth place is known — and hence significant.

.136 has *three* significant figures.

.1360 has *four* significant figures; the zero fixes the fourth place.

.00136 has *three* significant figures; the first two zeros merely locate the decimal point.

2.00136 has *six* significant figures; the integer, 2, makes the two zeros to the right of the decimal point significant.

3. Exact and Approximate Numbers

It is necessary in calculation to make a distinction between *exact* and *approximate* numbers. An exact number is one which is found by counting: ten children, 150 test scores, twenty desks are examples. Approximate numbers result from the measurement of variable quantities. Test scores and other measures, for example, are approximate since they are represented by intervals and not exact points on some scale. Thus a score of 61 may be any value from 60.5 up to 61.5 and a measured height of 47.5 inches may be any value from 47.45 up to 47.55 inches (see p. 3). Calculations with exact numbers may, in general, be carried to as many decimals as we please, since we may assume as many significant figures as we wish. For example, 110 test scores, which means that exactly 110 subjects were tested, could be written $N = 110.000 \ldots$ to n significant figures. Calculations based upon approximate numbers depend upon, and are limited by, the number of significant figures

in the numbers which enter into the calculations. This will be clearer in the following "rules":

4. Rules for Computation

(1) Accuracy of a Product

(a) The number of significant figures in the product of two or more approximate numbers will equal the number of significant figures in that one of the numbers which is the least accurate, i.e., which contains the smallest number of significant figures. To illustrate:

$125.5 \times 7.0 =$ 880, not 878.5, because 7.0, the less accurate of the two numbers, contains only two significant figures. The number 125.5 contains four significant figures.

$125.5 \times 7.000 =$ 878.5. Both numbers now contain four significant figures; hence their product also contains four significant figures.

(b) When multiplying an exact number by an approximate number, the number of significant figures in the product is determined by the number of significant figures in the approximate number. To illustrate:

If each of twelve children (twelve is an exact number) has an M.A. of eight years (eight is an approximate number) the product 12×8 must be written either as 90 or 100, since the approximate number has only *one* significant digit. If, however, each M.A. of eight years can be written as 8.0, the product 12×8.0 can be written as 96, since 8.0 contains *two* significant digits.

(2) Accuracy of a Quotient

(a) When dividing one approximate number by another approximate number, the significant figures in the quotient will equal the significant figures in that one of the two numbers (dividend or divisor) which is less accurate, i.e., which has the smaller number of significant digits. Illustrations:

$\frac{9.27}{41}$ should be written .23, not .22609, since 41 (the less accurate number) contains only two significant figures.

$\frac{16}{4724}$ should be written .0034, not .0033869, since 16 (the less accurate number) has two significant figures.

(b) In dividing an approximate number by an exact number, the number of significant figures in the quotient will equal the number of significant figures in the approximate number. Illustrations:

$\frac{9.27}{41}$ should be written .226, since 9.27, the approximate number, has three significant figures. The number 41 is an exact number.

$\frac{8541}{50}$ should be written 170.8, not 170.82, since 8541, the approximate number, contains only four significant figures.

(c) In dealing with exact numbers, quotients may be written to as many decimals as one wishes.

(3) Accuracy of a Root or Power

(a) The square root of an approximate number can contain no more significant figures than there are in the number itself. The number of significant figures retained in a square root is usually less than (often one-half) the number of significant figures in the number. For example, $\sqrt{159.5600}$ is usually written 12.63, and not 12.63176, although the original number, 159.5600, contains seven significant figures.

(b) The square, or higher power, of an approximate number contains as many significant figures as there are in the original number (and no more). For example, $(.034)^2 = .0012$ (two significant figures) and not .001156 (four significant figures).

(c) Roots and powers of exact numbers may be taken to as many decimal places as one wishes.

(4) Accuracy of a Sum or Difference

The number of decimal places to be retained in a sum or difference should be no greater than the *number of decimals* in the least accurate of the numbers added or subtracted. Illustrations:

$362.2 + 18.225 + 5.3062 =$ 385.7 not 385.7312, since the least accurate number (362.2) contains only one decimal.

$362.2 - 18.245 =$ 344.0, not 343.955, since the less accurate number (362.2) contains only one decimal.

PROBLEMS

1. Indicate which of the following variables fall into continuous and which into discrete series: (*a*) time; (*b*) salaries in a large business firm; (*c*) sizes of elementary school classes; (*d*) age; (*e*) census data; (*f*) distance traveled by car; (*g*) football scores; (*h*) weight; (*i*) numbers of pages in 100 books; (*j*) mental ages.
2. Write the exact upper and lower limits of the following scores in accordance with the two definitions of a score in continuous series, given on pages 3 and 4:

62	175	1
8	312	87

3. Suppose that sets of scores have the ranges given below. Indicate how large an interval, and how many intervals, you would suggest for use in drawing up a frequency distribution of each set.

Range	Size of Interval	Number of Intervals
16 to 87		
0 to 46		
110 to 212		
63 to 151		
4 to 12		

4. In each of the following write (*a*) the exact lower and upper limits of the class-intervals (following the first definition of a score, given on page 3), and (*b*) the midpoint of each interval.

45–47	162.5–167.5	63–67	0–9
1–4	80 up to 90	16–17	25–28

5. (*a*) Tabulate the following twenty-five scores into two frequency distributions, using (1) an interval of three, and (2) an interval of five units. Let the first interval begin with the score of 60.

72	75	77	67	72
81	78	65	86	73
67	82	76	76	70
83	71	63	72	72
61	67	84	69	64

(*b*) The following 100 scores were made on the Thorndike Intelligence Examination for High School Graduates by applicants

for admission to college. Tabulate these scores into three frequency distributions, using class-intervals of three, five, and ten units. Let the first interval begin with the score 45.

63	78	76	58	95
78	86	80	96	94
46	78	92	86	88
82	101	102	70	50
74	65	73	72	91
103	90	87	74	83
78	75	70	84	98
86	73	85	99	93
103	90	79	81	83
87	86	93	89	76
73	86	82	71	94
95	84	90	73	75
82	86	83	63	56
89	76	81	105	73
73	75	85	74	95
92	83	72	98	110
85	103	81	78	98
80	86	96	78	71
81	84	81	83	92
90	85	85	96	72

6. (*a*) Plot frequency polygons for the two distributions of twenty-five scores found in 5(*a*), using intervals of three and of five score units. Smooth both distributions (see p. 16) and plot the smoothed *f*'s and the original scores on the same axes.
   (*b*) Plot a frequency polygon of the 100 scores in 5(*b*) using an interval of ten score units. Superimpose a histogram upon the frequency polygon.
   (*c*) On the same axes, plot a frequency polygon and histogram of the 100 Thorndike scores using an interval of five score units. Smooth the frequency polygon and plot on the same diagram.

7. Reduce the distributions A and B below to percentage frequencies and plot them as frequency polygons on the same axes. Is your understanding of the achievement of these groups advanced by this treatment of the data?

Scores	Group A	Group B
52–55	1	8
48–51	0	5
44–47	5	12
40–43	10	58
36–39	20	40
32–35	12	22
28–31	8	10
24–27	2	15
20–23	3	5
16–19	4	0
	65	175

8. (*a*) Round off the following numbers to two decimals:

3.5872	74.168	126.83500
46.9223	25.193	81.72558

(*b*) How many significant figures in each of the following:

.00046	91.00	1.03
46.02	18.365	15.0048

(*c*) Write the answers to the following:

$127.4 \times .0036 =$ (both numbers approximate)

$200.0 \div 5.63 =$ " " "

$62 \times .053 =$ (first number exact, second approximate)

$364.2 + 61.596 =$

$364.2 - 61.596 =$

$\sqrt{47.86} =$

$(18.6)^2 =$

Answers

2. 61.5 to 62.5 and 62.0 to 63.0; 174.5 to 175.5 and 175.0 to 176.0;
7.5 to 8.5 and 8.0 to 9.0; 311.5 to 312.5 and 312.0 to 313.0;
.5 to 1.5 and 1.0 to 2.0
86.5 to 87.5 and 87.0 to 88.0

3.

Size of Interval	No. of Intervals
5	15
3 or 4 or 5	16 or 12 or 10
10	11
5 or 10	18 or 9
1	9

4.

	Midpoint
44.5 to 47.5	46.0
.5 to 4.5	2.5
162.5 to 167.5	165.0
79.5 to 89.5	84.5
62.5 to 67.5	65.0
15.5 to 17.5	16.5
− .5 to 9.5	4.5
24.5 to 28.5	26.5

8.

(*a*)	3.59	74.17	126.83
	46.92	25.19	81.73
(*b*)	2	4	3
	4	5	6
(*c*)	.46		
	35.5		
	3.3		
	425.8		
	302.6		
	6.918 or 6.92		
	346		

CHAPTER II

MEASURES OF CENTRAL TENDENCY

WHEN scores or other measures have been tabulated into a frequency distribution, as shown in Chapter I, usually the next task is to calculate one or more measures of *central tendency.* The value of a measure of central tendency is twofold. *First,* it is a single measure which represents *all* of the scores made by the group, and as such gives a concise description of the performance of the group as a whole; and *second,* it enables us to compare two or more groups in terms of typical performance. There are three "averages" or measures of central tendency in common use, (1) the *arithmetic mean,* (2) the *median,* and (3) the *mode.* Popularly, the *average* is the term used for the arithmetic mean. In statistical work, however, the term average is often used as a general expression to cover any measure of central tendency.

I. CALCULATION OF MEASURES OF CENTRAL TENDENCY

1. The Arithmetic Mean or "Average" (M)

(1) Calculation of the Mean When Data Are Ungrouped

The arithmetic mean or simply the mean is the best known measure of central tendency. It may be defined as the sum of the separate scores or other measures divided by their number. To illustrate: if a man earns \$3, \$4, \$3.50, \$5, and \$4.50 on five successive days his mean daily wage (\$4.00) is obtained by dividing the sum of his daily earnings by the number of days he has worked. The formula for the arithmetic mean (M) of a series of ungrouped measures is

$$M = \frac{\Sigma X}{N} \tag{1}$$

(*arithmetic mean calculated from ungrouped data*)

in which N is the number of measures in the series, X stands for a score or other measure, and the symbol Σ means "sum of," here sum of scores.

(2) Calculation of the Mean from Data Grouped into a Frequency Distribution

When measures have been grouped into a frequency distribution, the arithmetic mean is calculated by a slightly different method from the one given above. The two illustrations given in Table 5 will make the differences clear. The first example shows the calculation of the mean of the fifty Army Alpha scores which were tabulated into a frequency distribution in Table 1. First calculate the fX column by multiplying the midpoint (X) of each interval by the number of scores (f) on it; the mean (170.80) is then simply the sum of the fX (namely, 8540) * divided by N (50). The use of the midpoint for all of the scores within an interval is made necessary by the fact that scores grouped into intervals lose their identity and must thereafter be represented by the midpoint of that particular interval in which they fall. Hence, we multiply or "weight" the midpoint of each interval by the frequency upon that interval; add the fX and divide by N to obtain the mean. The formula may be written

$$M = \frac{\Sigma fX}{N} \tag{2}$$

(arithmetic mean calculated from scores grouped into a frequency distribution)

The second example in Table 5 is another illustration of the calculation of the mean from grouped data. This frequency distribution represents 200 scores made by a group of adults upon a cancellation test. Scores have been classified by method (B), page 7, into nine class-intervals; and since the

* The sum 8540 may be written 8540.000 . . . (i.e., to any number of significant figures) since each midpoint value (x) is an exact point within a score interval, and the f's are exact numbers. The mean (170.80) has been carried only to two decimals — the usual standard of accuracy for measures of central tendency.

intervals are four units, the midpoints are found by adding one-half of four to the lower limit of each. For example, in the first interval, $103.5 + 2.0 = 105.5$. The fX column totals 23,888.0; and N equals 200. Hence, applying formula (2), the arithmetic mean is found to be 119.44 (to two decimals).

In both of the illustrations in Table 5, the M of the scores made by the members of a *group* was found. We may, however, use either formula (1) or formula (2) to calculate the M of a number of measurements made upon the same individual. If an individual's reaction time to light is measured 100 times, and the measures tabulated into a frequency distribution, the M is found in exactly the same way in which we compute the "average" reaction time to light of 100 *different* observers.

2. The Median (Mdn) *

(1) Calculation of the Median When Data Are Ungrouped

When ungrouped scores or other measures are arranged in order of size, the median is the *midpoint* in the series. Two situations arise in the computation of the median from ungrouped data: (*a*) when N is *odd*, and (*b*) when N is *even*. To consider, first, the case where N is odd, suppose we have the following integral "mental ages" — 7, 10, 8, 12, 9, 11, 7, calculated from seven performance tests. If we arrange these seven scores in order of size

$$7 \quad 7 \quad 8 \quad (9) \quad 10 \quad 11 \quad 12$$

the median is 9.0 since 9.0 is the midpoint of that score which lies midway in the series. Calculation is as follows: There are three scores above, and three below 9, and since a score of 9 covers the interval 8.5 to 9.5, its midpoint is 9.0. This is the median.

Now if we drop the first score of 7 our series contains six scores

$$7 \quad 8 \quad 9 \overset{9.5}{\uparrow} 10 \quad 11 \quad 12$$

and the median is 9.5. Counting three scores in from the beginning of the series, we *complete* score 9 (which is 8.5 to 9.5)

* The median is also designated as *Md*.

TABLE 5

THE CALCULATION OF THE MEAN, MEDIAN, AND CRUDE MODE FROM DATA GROUPED INTO A FREQUENCY DISTRIBUTION

1. Data from Table 1, fifty Army Alpha scores
Class-interval = 5

Class-Intervals Scores	Midpoint X	f		fX
195–199	197	1		197
190–194	192	2		384
185–189	187	4	↓	748
180–184	182	5		910
175–179	177	8	20	1416
170–174	172	10		1720
165–169	167	6	20	1002
160–164	162	4	↑	648
155–159	157	4		628
150–154	152	2		304
145–149	147	3		441
140–144	142	1		142
		$N = 50$		8540
		$N/2 = 25$		

(1) Mean $= \frac{\Sigma fX}{N} = \frac{8540}{50} = 170.80$

(2) Median $= 169.5 + \frac{5}{10} \times 5 = 172.00$

(3) Crude Mode falls on class-interval 170–174 or at 172.00

2. Scores made by 200 adults upon a cancellation test
Class-interval = 4

Class-Intervals Scores	Midpoint X	f		fX
135.5 to 139.5	137.5	3		412.5
131.5 to 135.5	133.5	5		667.5
127.5 to 131.5	129.5	16	↓	2072.0
123.5 to 127.5	125.5	23		2886.5
119.5 to 123.5	121.5	52	99	6318.0
115.5 to 119.5	117.5	49		5757.5
111.5 to 115.5	113.5	27	52	3064.5
107.5 to 111.5	109.5	18	↑	1971.0
103.5 to 107.5	105.5	7		738.5
		$N = 200$		23888.0
		$N/2 = 100$		

(1) Mean $= \frac{\Sigma fX}{N} = \frac{23,888.0}{200} = 119.44$

(2) Median $= 115.5 + \frac{48}{49} \times 4 = 119.42$

(3) Crude Mode falls on class-interval 119.5 to 123.5 or at 121.50

to reach 9.5, the *upper* limit of score 9. In like manner, counting three scores in from the end of the series, we move *through* score 10 (10.5 to 9.5) reaching 9.5, the *lower* limit of score 10.

A formula for finding the median of a series of ungrouped scores is

$$\text{Median} = \text{the } \frac{(N+1)}{2}\text{th measure in order of size} \qquad (3)$$

(*median from ungrouped data*)

In our first illustration above, the median is on the $\frac{(7+1)}{2}$ or fourth score counting in from either end of the series, that is, 9.0 (midpoint 8.5 to 9.5). In our second illustration, the median is on the $\frac{(6+1)}{2}$ or 3.5th score in order of size, that is, 9.5 (upper limit of score 9, or lower limit of score 10).

(2) Calculation of the Median When Data Are Grouped into a Frequency Distribution

When scores in a continuous series are grouped into a frequency distribution, the median by definition is the 50% *point* in the distribution. To locate the median, therefore, we take 50% (i.e., $N/2$) of our scores, and count into the distribution until the 50% point is reached. The method is illustrated in the two examples in Table 5. Since there are fifty scores in the first distribution, $N/2 = 25$, and the median is that point in our distribution of Army Alpha scores which has twenty-five scores on each side of it. Beginning at the small-score end of the distribution, and adding up the scores in order, we find that intervals 140–144 to 165–169, inclusive, contain just 20 f's — five scores short of the twenty-five necessary to locate the median. The next interval, 170–174, contains ten scores assumed to be spread evenly over the interval (p. 8). In order to get the five extra scores needed to make exactly twenty-five, we take $5/10 \times 5$ (the length of the interval) and add this increment (2.5) to 169.5, the beginning of the interval 170–174.

This puts the *Mdn* at $169.5 + 2.5$ or at 172.0. The reader should note carefully that the median like the mean is a *point* and not a *score*.

A second illustration of the calculation of the median from data grouped into a frequency distribution is given in Table 5 (2). There are 200 scores in this distribution; hence, $N/2 = 100$, and the median must lie at a point 100 scores distant from either end of the distribution. If we begin at the small-score end of the distribution (103.5 to 107.5) and add the scores in order, fifty-two scores take us *through* the interval 111.5 to 115.5. The 49 scores on the next interval (115.5 to 119.5) plus the fifty-two already counted off total 101 — *one* score too many to give us 100, the point at which the median falls. To get the forty-eight scores needed to make *exactly* 100 we must take $48/49 \times 4$ (the length of the interval) and add this amount (3.92) to 115.5, the beginning of interval 115.5 to 119.5. This procedure takes us exactly 100 scores into the distribution, and locates the median at 119.42.

A formula for calculating the *Mdn* when the data have been classified into a frequency distribution is

$$Mdn = l + \left(\frac{\frac{N}{2} - F}{f_m}\right)i \tag{4}$$

(*median computed from data grouped into a frequency distribution*)
where

l = lower limit of the class-interval upon which the median lies

$\frac{N}{2}$ = one-half the total number of scores

F = sum of the scores on all intervals *below* l

f_m = frequency (number of scores) *within* the interval upon which the median falls

i = length of the class-interval

To illustrate the use of formula (4), consider the first example in Table 5. Here $l = 169.5$, $N/2 = 25$, $F = 20$, $f_m = 10$, and

$i = 5$. Hence, the median falls at $169.5 + \frac{(25 - 20)}{10} \times 5$ or at 172.0. In the second example, $l = 115.5$, $N/2 = 100$, $F = 52$, $f_m = 49$, and $i = 4$. The median, therefore, is $115.5 + \frac{(100 - 52)}{49} \times 4$ or 119.42.

The steps involved in computing the *Mdn* from data tabulated into a frequency distribution may be summarized as follows:

(1) Find $N/2$, that is, one-half of the cases in the distribution.
(2) Begin at the small-score end of the distribution and count off the scores in order up to the lower limit (l) of the interval which contains the median. The sum of these scores is F.
(3) Compute the number of scores necessary to fill out $N/2$, i.e., compute $N/2 - F$. Divide this quantity by the frequency (f_m) on the interval which contains the median; and multiply the result by the size of the class-interval (i).
(4) Add the amount obtained by the calculations in (3) to the lower limit (l) of the interval which contains the *Mdn*. This will give the median of the distribution.

The median may also be computed by adding up one-half of the scores from the top down in a frequency distribution. The procedure is the same through step (3) in the summary above. When we count down from the top of the distribution, however, the quantity found in step (3) must be *subtracted* from the *upper* limit of the interval containing the median. To illustrate with the data of Table 5 (1), counting down in the f-column, twenty scores *complete* interval 175–179, and we reach 174.5, the upper limit of the interval 170–174. Five scores of the ten on this interval are needed to make twenty-five ($N/2$). Hence we have $174.5 - \frac{5}{10} \times 5 = 172.0$, which checks our first calculation of the median. In Table 5 (2), the median found by counting down is $119.5 - \frac{4}{49} \times 4$ or 119.42.

(3) Calculation of the *Mdn* When (*a*) the Frequency Distribution Contains Gaps; and When (*b*) the First or Last Interval Has Indeterminate Limits

(*a*) Difficulty arises when it becomes necessary to calculate the median from a distribution in which there are gaps or zero frequency upon one or more intervals. The method to be followed in such cases is shown in Table 6. Since $N = 10$, and $N/2 = 5$, we count *up* the frequency column five scores through 6–7. Ordinarily, this would put the median at 7.5, the lower limit of interval 8–9. If we check this median, however, by counting *down* the frequency column five scores, the median falls at 11.5, the lower limit of 12–13. Obviously, the discrepancy between these two values of the median is due to the two intervals 8–9 and 10–11 (each of which has zero frequency) which lie between 6–7 and 12–13. In order to have the median come out at the same point, whether computed from the top or the bottom of the frequency distribution, the procedure usually followed in cases like this is to have interval 6–7 *include* 8–9, thus becoming 6–9; and to have interval 12–13 *include* 10–11, becoming 10–13. Lengthening these intervals

TABLE 6

COMPUTATION OF THE MEDIAN WHEN THERE ARE GAPS IN THE DISTRIBUTION

Class-Intervals Scores	f	
20–21	2	
18–19	1	
16–17	0	
14–15	0	
12–13	2	10–13
10–11	0	
8–9	0	6–9
6–7	2	
4–5	1	
2–3	1	
0–1	1	
$N =$	10	
$N/2 =$	5	

$$Mdn = 9.5 + \frac{0}{2} \times 2 = 9.5$$

from two to four units eliminates the zero frequency on the adjacent intervals by spreading the numerical frequency over them. If now we count off five scores, going *up* the frequency column through 6–9, the median falls at 9.5, the upper limit of this interval. Also, counting *down* the frequency column five scores, we arrive at a median value of 9.5, the *upper* limit of 6–9, *or* the *lower* limit of 10–13. Computation from the two ends of the series now gives consistent results — the median is 9.5 in both instances.

(*b*) When scores scatter widely, the last class-interval in a frequency distribution may be designated as "80 and above" or simply as 80 +. This means that *all* scores above 80 are thrown into this interval, the upper limit of which is indeterminate. The same lumping together of scores may also occur at the beginning of the distribution, when the first interval, for example, is designated "20 and below" or 20 −. The lower limit of the beginning class-interval is now indeterminate. In irregular distributions like these, the median is readily computed since each score is simply counted as one frequency whether accurately classified or not. But it is impossible to calculate the mean exactly when the midpoint of one or more intervals is unknown. The mean depends upon the absolute size of the scores (or their midpoints) and is directly affected by indeterminate interval limits.

6. The Mode

In a simple ungrouped series of measures the "crude" or "empirical" mode is that single measure or score which occurs most frequently. For example, in the series 10, 11, 11, 12, 12, 13, 13, 13, 14, 14, the most often recurring measure, namely 13, is the crude or empirical mode. When data are grouped into a frequency distribution, the crude mode is usually taken to be the midpoint of that interval which contains the largest frequency. In example 1, Table 5, the interval 170–174 contains the largest frequency and hence 172.0, its midpoint, is the crude mode. In example 2, Table 5, the largest frequency

falls on 119.5 to 123.5 and the crude mode is at 121.5, the midpoint.

When calculating the mode from a frequency distribution, we distinguish between the "true" mode and the crude mode. The true mode is the point (or "peak") of greatest concentration in the distribution; that is, the point at which more measures fall than at any other point. When the scale is divided into finely graduated units, when scores are recorded exactly, and when N is large, the crude mode closely approaches the true mode. Ordinarily, however, the crude mode is only approximately equal to the true mode. A formula for approximating the true mode, when the frequency distribution is symmetrical, or at least not badly skewed (p. 119) is

$$Mode = 3\ Mdn - 2\ Mean \tag{5}$$

(approximation to the true mode calculated from a frequency distribution)

If we apply this formula to the data in Table 5, the mode is 174.40 for the first distribution, and 119.38 for the second. The first mode is somewhat larger and the second slightly smaller than the crude modes obtained from the same distributions.

The crude mode is often an unstable measure of central tendency. This instability is not, however, so serious a drawback as might seem at first glance. The crude mode is usually employed as a simple, inspectional "average," to indicate in a rough way the center of concentration in the distribution; and for this purpose it need not be calculated as exactly as the median and mean.

II. Calculation of the Mean by the "Assumed Mean" or Short Method

In Table 5 the mean was calculated by multiplying the midpoint (X) of each interval by the frequency (number of scores) on the interval, summing up these values (the fX column) and

dividing by N, the number of scores. This straightforward method (called the Long Method) gives accurate results but often requires the handling of large numbers and entails tedious calculation. Because of this, the "Assumed Mean" method, or simply the Short Method, has been devised for computing the mean. The Short Method does not apply to the calculation of the median or the mode. These measures are always found by the methods previously described.

The most important fact to remember in calculating the mean by the Short Method is that we "guess" or "assume" a mean at the outset, and later apply a correction to this assumed value (AM) in order to obtain the actual mean (M) (see Table 7). There is no set rule for assuming a mean.* The best plan is to take the midpoint of an interval somewhere near the *center* of the distribution; and if possible the midpoint of that interval which contains the largest *frequency*. In Table 7, the largest f is on interval 170–174, which also happens to be almost the center of the distribution. Hence the AM is taken at 172.0, the middle of this interval. When the question of the AM is settled, we determine the correction which must be applied to the AM in order to get M. Steps are as follows:

(1) First, we fill in the x' column,† column (4). Here are entered the deviations of the midpoints of the different steps measured from the AM in *units of class-interval*. Thus 177, the midpoint of 175–179, deviates from 172, the AM, by *one* interval; and a "1" is placed in the x' column opposite 177. In like manner, 182 deviates *two* intervals from 172; and a "2" goes in the x' column opposite 182. Reading on up the x' column from 172, we find the succeeding entries to be 3, 4, and 5. The last entry, 5, is the interval-deviation of 197 from 172; the actual score-deviation, of course, is 25.

* The method outlined here gives consistent results no matter where the mean is tentatively placed or assumed.

† x' is regularly used to denote the deviation of a score X from the assumed mean (AM); x is the deviation of a score X from the actual mean (M) of the distribution.

Returning to 172, we find that the x' of this midpoint measured from the AM (from itself) is zero; hence a zero is placed in the x' column opposite 170–174. Below 172, all of the x' entries are negative, since all of the midpoints are less than 172, the AM. So the x' of 167 from 172 is -1 interval; and the x' of 162 from 172 is -2 intervals. The other x''s are -3, -4, -5, and -6 intervals.

(2) The x' column completed, we compute the fx' column, column (5). The fx' entries are found in exactly the same way as are the fX in Table 5, page 35. Each x' in column (4) is multiplied or "weighted" by the appropriate f in column (3). Note again that in the Short Method we multiply each x' by its deviation from the AM in *units of class-interval,* instead of by its actual deviation from the mean of the distribution. For this reason, the computation of the fx' column is much more simple than is the calculation of the fX column by the method given on page 33. All of the fx' on intervals *above* (greater than)

TABLE 7

THE CALCULATION OF THE MEAN BY THE SHORT METHOD
(Data from Table 1, fifty Army Alpha scores)

(1) Class-Intervals Scores	(2) Midpoint X	(3) f	(4) x'	(5) fx'
195–199	197	1	5	5
190–194	192	2	4	8
185–189	187	4	3	12
180–184	182	5	2	10
175–179	177	8	1	8
170–174	172	10	0	+43
165–169	167	6	−1	−6
160–164	162	4	−2	−8
155–159	157	4	−3	−12
150–154	152	2	−4	−8
145–149	147	3	−5	−15
140–144	142	1	−6	−6
		$N = 50$		−55

$AM = 172.00$
$ci = -1.20$
$M = 170.80$

$c = -\frac{12}{50} = -.240$
$i = 5$
$ci = -.240 \times 5 = -1.20$

the AM are *positive;* and all fx' on intervals *below* (smaller than) the AM are *negative,* since the signs of the fx' depend upon the signs of the x'.

(3) From the fx' column the correction is obtained as follows: The sum of the positive values in the fx' column is 43; and the sum of the negative values in the fx' column is -55. There are, therefore, 12 more *minus* fx' values than *plus* (the algebraic *sum* is -12); and -12 divided by 50 (N) gives $-.240$ which is the correction (c) in *units of class-interval.* If we multiply c ($-.240$) by i, the length of the interval (here 5), the result is ci (-1.20) the score correction, or the correction in *score units.* When -1.20 is added to 172.00, the AM, the result is the actual mean, 170.80.

The process of calculating the mean by the Short Method may be summarized as follows:

(1) Tabulate the scores or measures into a frequency distribution.

(2) "Assume" a mean as near the center of the distribution as possible, and preferably on the interval containing the largest frequency.

(3) Find the deviation of the midpoint of each class-interval from the AM in units of interval.

(4) Multiply or weight each deviation (x') by its appropriate f — the f opposite it.

(5) Find the algebraic sum of the plus and minus fx' and divide this sum by N, the number of cases. This gives c, the correction in units of class-interval.

(6) Multiply c by the interval length (i) to get ci, the score correction.

(7) Add ci algebraically to the AM to get the actual mean. Sometimes ci will be positive and sometimes negative, depending upon where the mean has been assumed. The method works equally well in either case.

III. When to Use the Various Measures of Central Tendency

The beginning student of statistics is often puzzled to know which measure of central tendency to use in a given problem. The following summary will serve as a convenient guide for most statistical work.

1. *Use the mean*
   (1) When each score or measure should have *equal* weight in determining the central tendency. Since the mean is the sum of the scores divided by their number, each score has equal weight in its determination.
   (2) When the measure of central tendency having the highest reliability is desired. (p. 193)
   (3) When standard deviations and product-moment coefficients of correlation are to be subsequently computed. (p. 282)
   4) When 1st or last class interval has indeterminate
2. *Use the median*
   (1) When a quick and easily computed measure of central tendency is wanted.
   (2) When there are extreme measures which would affect the mean disproportionately (p. 39).
   (3) When it is desired that certain scores should influence the central tendency but all that is known about them is that they are above or below the median (p. 40).
3. *Use the mode*
   (1) When the most often recurring score is sought.
   (2) When a quick approximate measure of concentration is all that is wanted.

PROBLEMS

1. Calculate the mean, median, and mode for the following frequency distributions. Use the Short Method in computing the mean.

(1) Scores	f
70–71	2
68–69	2
66–67	3
64–65	4
62–63	6
60–61	7
58–59	5
56–57	4
54–55	2
52–53	3
50–51	1
	$N = 39$

(2) Scores	f
90–94	2
85–89	2
80–84	4
75–79	8
70–74	6
65–69	11
60–64	9
55–59	7
50–54	5
45–49	0
40–44	2
	$N = 56$

(3) Scores	f
120–122	2
117–119	2
114–116	2
111–113	4
108–110	5
105–107	9
102–104	6
99–101	3
96–98	4
93–95	2
90–92	1
	$N = 40$

(4) Scores	f
100–109	5
90–99	9
80–89	14
70–79	19
60–69	21
50–59	30
40–49	25
30–39	15
20–29	10
10–19	8
0–9	6
	$N = 162$

2. Compute the mean and the median for each of the two distributions in problem 5(a), page 28, tabulated in three- and five-unit intervals. Compare the two means and the two medians, and explain any discrepancy found. (Let the first interval in the first distribution be 61–63; the first interval in the second distribution, 60–64.)

3. (*a*) Compute the median of the following sixteen scores:

Scores	f
20 to 22	2
18 to 20	2
16 to 18	0
14 to 16	4
12 to 14	0
10 to 12	0
8 to 10	4
6 to 8	0
4 to 6	0
2 to 4	0
0 to 2	4
	$N = 16$

(*b*) In a group of fifty children, the eight children who took longer than five minutes to complete a performance test were marked D.N.C. (did not complete). In computing a measure of central tendency for this distribution of scores, what measure would you use, and why?

(*c*) Find the medians of the following arrays of ungrouped scores:
(1) 21, 24, 27, 29, 29, 30, 32, 33, 35, 38, 42, 45.
(2) 54, 59, 64, 67, 70, 72, 73, 75, 78, 83, 90.
(3) 7, 8, 9, 9, 10, 11.

4. The time by your watch is 10:31 o'clock. In checking with two friends, you find that their watches give the time as 10:25 and 10:34. Assuming that the three watches are equally good timepieces, what do you think is probably the "correct time"?

5. What is meant popularly by the "law of averages"?

6. (*a*) When one uses the term "in the mode" does he have reference to the mode of a distribution?

(*b*) What is approximately the modal time for each of the following meals: breakfast, lunch, dinner. Explain your answers.

(*c*) Why is the median usually the best measure of the typical contribution in a church collection?

Answers

1. (1) Mean = 60.76
   Median = 60.79
   Mode = 60.85

   (2) Mean = 67.36
   Median = 66.77
   Mode = 65.59

   (3) Mean = 106.00
   Median = 105.83
   Mode = 105.49

   (4) Mean = 55.43
   Median = 55.17
   Mode = 54.65

2. Class-interval = 3
   Mean = 72.92
   Median = 71.75

   Class-interval = 5
   Mean = 73.00
   Median = 72.71

3. (*a*) Median = 11.5

   (*c*) (1) Median = 31.0
   (2) Median = 72.0
   (3) Median = 9.0

4. Mean is 10:30.

CHAPTER III

MEASURES OF VARIABILITY

In Chapter II the calculation of three measures of central tendency — measures typical or representative of a set of scores as a whole — was described. Ordinarily, the next step is to find some measure of the *variability* of our scores, i.e., of the "scatter" or "spread" of the separate scores or measures around their central tendency. It will be the task of this chapter to show how measures of variability may be computed.

The usefulness of a measure of variability can be seen from a simple example. Suppose a test of controlled association has been administered to a group of fifty boys and to a group of fifty girls. The mean scores are, boys, 34.6 seconds, and girls, 34.5 seconds. So far as the means go there is no difference in the performance of the two groups. But suppose the boys' scores are found to range from 15 to 51 seconds and the girls' scores from 19 to 45 seconds. This difference in range shows that in a general way the boys "cover more territory," are more *variable*, than the girls; and this greater variability may be of more interest than the lack of a difference in the means. If a group is *homogeneous*, that is, made up of individuals of nearly the same ability, most of the scores will fall around the same point on the scale, the range will be relatively short, and the variability will be small. But if the group contains individuals of widely differing capacities, scores will be strung out from high to low, the range will be relatively wide, and the variability large.

This situation is represented graphically in Figure 7, which shows two frequency distributions of the same area (N) and same mean (50) but of very different variability. Group A ranges from 20 to 80, and Group B from 40 to 60. Group A is three times as variable as Group B — spreads over three times the distance on the scale of scores — though both distributions have the same central tendency.

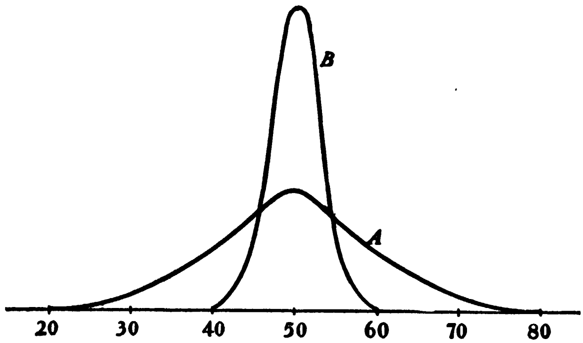


FIG. 7. Two Distributions of the Same Area (N) and Mean (50) but of Very Different Variability.

Four measures have been devised to indicate the variability or dispersion within a set of measures. These are (1) the *range*, (2) the *quartile deviation* or Q, (3) the *mean deviation* or MD, and (4) the *standard deviation* or SD.

I. CALCULATION OF MEASURES OF VARIABILITY

1. The Range

In grouping the scores in Table 1 into a frequency distribution (p. 6) we have already had occasion to use the *range*. It may be redefined simply as the interval between the largest and the smallest scores. In the illustration above, the range of boys' scores was 51–15 or 36 seconds and the range of girls' scores 45–19 or 26 seconds. The range is the most general measure of spread or scatter, and is computed when we wish to make a rough comparison of two or more groups for variability. Since the range takes account of the extremes of the series only it is unreliable when N is small or when many or large gaps (i.e. zero f's) occur in the frequency distribution.

2. The Quartile Deviation or Q

The *quartile deviation* or Q is one-half of the distance between the 75th and 25th percentiles in a frequency distribution.

The 25th percentile, called Q_1, is the *first* quarter or *quartile* on the scale of scores, the point below which lie 25% of the scores. The 75th percentile, or Q_3, is the *third* quarter or *quartile* on the score-scale, the point below which lie 75% of the scores.*

To find Q, we must first calculate the 75th and 25th percentiles. These values are found by exactly the same method employed in calculating the median. To find Q_1, count off 25% of the scores from the beginning of the distribution (low end); and to find Q_3 count off 75% of the scores from the low end of the distribution, or 25% from the high end.

Table 8 illustrates the calculation of Q for the distribution of fifty Alpha scores tabulated in Table 1. First, to find Q_1, count off 1/4 of N (12.5) from the *low*-score end of the distribution. When the scores (f) are added in order, the first four class-intervals (140–144 to 155–159, inclusive) are found to contain 10 scores. The next interval, 160–164, contains four scores, assumed to be spread evenly over the interval. Since we need only 2.5 additional scores to make up the necessary 12.5, take $2.5/4 \times 5$ (the interval) and add this amount, 3.13, to 159.50, the beginning of the interval which contains Q_1. This calculation locates Q_1 at 162.63 (see Table 8).

Q_3 is found in the same way by counting off 3/4 of N (37.5) from the small-score end of the distribution. The f's on 140–144 to 170–174, inclusive, added in order, total 30. The next interval, 175–179, contains eight scores. To make up the necessary 37.5, therefore, take $7.5/8 \times 5$ (interval) and add this amount (4.69) to 174.50. This puts Q_3 at 179.19 (see Table 8).

When Q_1 and Q_3 are known, Q, the quartile deviation, is found from the formula

$$Q = \frac{Q_3 - Q_1}{2} \qquad (6)$$

(*quartile deviation calculated from grouped data*)

In the present problem, $Q = \dfrac{179.19 - 162.63}{2}$ or 8.28.

* It may be noted that the second quartile, Q_2, is the median.

A second illustration of the calculation of Q from a frequency distribution is given in Table 8, example 2. Since the N of this distribution is 200, 1/4 of N equals 50. The intervals 103.5 to 107.5 and 107.5 to 111.5 contain twenty-five scores; and the next interval, 111.5 to 115.5, contains twenty-seven scores, which makes a total of fifty-two — two more than the fifty wanted. To find the point reached by just fifty scores, take $25/27 \times 4$ (the interval) and add this amount (3.70) to 111.50, the lower limit of 111.5 to 115.5. This locates Q_1 at 115.20.

To find Q_3 count off 3/4 of N or 150 scores from the small-score end of the distribution. The first four intervals include 101 scores, and the next interval, 119.5 to 123.5, contains fifty-two scores. To fill out the required 150, take $49/52 \times 4$, the length of the interval, and add this increment (3.77) to 119.50, to locate Q_3 at 123.27. Substituting 115.20 for Q_1 and 123.27 for Q_3 in formula (6) we get a Q of 4.04.

TABLE 8

THE CALCULATION OF THE Q, MD, AND SD FROM DATA GROUPED INTO A FREQUENCY DISTRIBUTION

1. Data from Table 1, fifty Army Alpha scores

(1) Class-Intervals Scores	(2) Midpoint X	(3) f		(4) x	(5) fx	(6) fx^2
195–199	197	1		26.20	26.20	686.44
190–194	192	2		21.20	42.40	898.88
185–189	187	4		16.20	64.80	1049.76
180–184	182	5		11.20	56.00	627.20
175–179	177	8		6.20	49.60	307.52
170–174	172	10	30	1.20	12.00	14.40
165–169	167	6		− 3.80	− 22.80	86.64
160–164	162	4		− 8.80	− 35.20	309.76
155–159	157	4	10	− 13.80	− 55.20	761.76
150–154	152	2		− 18.80	− 37.60	706.88
145–149	147	3		− 23.80	− 71.40	1699.32
140–144	142	1		− 28.80	− 28.80	829.44
		N = 50			502.00	7978.00

TABLE 8 (continued)

Mean = 170.80 (Table 5, p. 35)

$\frac{N}{4} = 12.5$, and, $\qquad \frac{3N}{4} = 37.5$, and,

$Q_1 = 159.5 + \frac{2.5}{4} \times 5 = 162.63 \qquad Q_3 = 174.5 + \frac{7.5}{8} \times 5 = 179.19$

$$Q = \frac{Q_3 - Q_1}{2} = \frac{179.19 - 162.63}{2} = 8.28$$

$$MD = \frac{|\Sigma fx|}{N} = \frac{502.00}{50} = 10.04$$

$$SD = \sqrt{\frac{\Sigma fx^2}{N}} = \sqrt{\frac{7978.00}{50}} = 12.63$$

2. Data from Table 3, p. 14, 200 cancellation scores

(1) Class-Intervals Scores	(2) Midpoint X	(3) f		(4) d. x	(5) fd. fx	(6) fx^2
135.5 to 139.5	137.5	3		18.06	54.18	978.49
131.5 to 135.5	133.5	5		14.06	70.30	988.42
127.5 to 131.5	129.5	16		10.06	160.96	1619.26
123.5 to 127.5	125.5	23		6.06	139.38	844.64
119.5 to 123.5	121.5	52		2.06	107.12	220.67
115.5 to 119.5	117.5	49	101	− 1.94	− 95.06	184.42
111.5 to 115.5	113.5	27		− 5.94	− 160.38	952.66
107.5 to 111.5	109.5	18	25	− 9.94	− 178.92	1778.46
103.5 to 107.5	105.5	7		− 13.94	− 97.58	1360.27
		$N = 200$			1063.88	8927.29

Mean = 119.44 (Table 5)

$\frac{N}{4} = 50$, and, $\qquad \frac{3N}{4} = 150$, and,

$Q_1 = 111.5 + \frac{25}{27} \times 4 = 115.20 \qquad Q_3 = 119.5 + \frac{49}{52} \times 4 = 123.27$

$$Q = \frac{Q_3 - Q_1}{2} = \frac{123.27 - 115.20}{2} = 4.04$$

$$MD = \frac{|\Sigma fx|}{N} = \frac{1063.88}{200} = 5.32$$

$$SD = \sqrt{\frac{\Sigma fx^2}{N}} = \sqrt{\frac{8927.29}{200}} = 6.68$$

The quartiles Q_1 and Q_3 mark off the limits of the *middle 50%* of scores in the distribution and the distance between these points is called the *interquartile range*. Q is one-half the

range of the middle 50% or the *semi-interquartile range*. Since Q measures the average distance of the quartile points from the median, it is a good measure of score density around the middle of the distribution. If the scores of a distribution are packed closely together the quartiles will be near to one another and Q will be small; if the scores are widely scattered, the quartiles will be relatively far apart, and Q will be large (see Fig. 7, p. 50).

When the distribution is asymmetrical or "skewed," Q_1 and Q_3 are at unequal distances from the median, and the difference between $(Q_3 - Mdn)$ and $(Mdn - Q_1)$ gives a measure of the amount and direction of the skewness (p. 119). When the distribution is symmetrical or *normal*, Q marks off exactly the 25% of cases just above, and the 25% of cases just below, the median. The median then lies just halfway between the two quartiles Q_1 and Q_3. In a normal distribution Q is commonly known as the *PE* (*probable error*). The terms Q and *PE* are often used interchangeably, but it is best to restrict the use of the term *PE* to the measurement of reliability (p. 187).

Steps in calculating Q may be summarized as follows:

To find Q_1

(1) Divide N by 4.
(2) Begin at the low-score end of the distribution, and count off the scores up to the interval which contains Q_1.
(3) Divide the number of scores necessary to locate Q_1 (i.e., to complete $N/4$) by the frequency in the interval reached in (2) above, and multiply the result by the class-interval.
(4) Add the amount obtained in (3) to the lower limit of the class-interval within which Q_1 lies. This gives Q_1.

To find Q_3

(1) Find 3/4 of N.
(2) Begin at the low-score* end of the distribution, and count up the scores until the interval which contains Q_3 is reached.

* Q_3 may also be found by counting in 25% from the high-score end of the distribution. To avoid confusion, the method given above is recommended to the beginner.

(3) Divide the number of scores required to locate Q_3 by the frequency within the interval reached in (2) and multiply the result by the class-interval.

(4) Add the amount obtained in (3) to the lower limit of the class-interval within which Q_3 lies. This gives Q_3.

To find Q

Substitute Q_3 and Q_1 in formula (6).

3. The Mean Deviation or *MD*

(1) Calculation of *MD* from Ungrouped Data

The *mean deviation* or *MD* (also written *average deviation* or *AD* and *mean variation* or *MV*) is the mean of the deviations of all the separate measures in a series taken from their central tendency (usually the arithmetic mean; less frequently the median or mode). In averaging deviations to find the *MD*, no account is taken of signs, and all deviations whether positive or negative are treated as positive.

An example will make our definition clearer. If we have five scores, 6, 8, 10, 12, and 14, the mean is easily found to be 10. It is then a simple process to find the deviation of each measure from this mean by subtracting the mean from each measure. Thus 6, the first score, minus 10 equals -4; $8 - 10 = -2$; $10 - 10 = 0$; $12 - 10 = 2$; and $14 - 10 = 4$. The five deviations measured from the mean are -4, -2, 0, 2, and 4. If we add these deviations *without regard to signs* the sum is 12; and dividing 12 by 5 (N), we get 2.4 as the *mean* of the five deviations from their mean, or the *MD*. The formula for the *MD* when scores are ungrouped may be written

$$MD = \frac{\Sigma \mid x \mid}{N} \tag{7}$$

(*mean deviation for ungrouped measures*)

in which the $\Sigma \mid x \mid$ denotes the sum of the deviations from the mean and N is, as before, the number of cases or items. The bars || enclosing Σx mean that signs are disregarded. The

small letter x in the formula always represents the deviation of a score X from its mean M, i.e., $x = X - M$.

(2) Calculation of *MD* from Grouped Data

In Table 8 the calculation of the *MD* for scores grouped into a frequency distribution is illustrated by two problems. The mean of the fifty Army Alpha scores in problem 1 has already been found in Table 5 to be 170.80. To compute the *MD* of the scores in this distribution we must take our deviations (x's) around this mean. However, since the scores have been grouped into class-intervals, we are unable to get the deviation of each *separate score* from the mean. In lieu of separate score deviations, therefore, we take the deviation of the *midpoint* of each interval from the mean. The substitution of the midpoint for all of the scores within an interval is the only difference between the computation of x's from grouped and from ungrouped data. The x of 195–199, for example, is 26.20, found by subtracting 170.80 (the mean) from 197.00 (the midpoint of the interval). All of the x's are positive as far down as 170–174, as in each case the midpoint is numerically larger than the mean. From the interval 165–169 on down to the beginning of the series, the x's are negative, as the midpoints of these intervals are all smaller than 170.80. Thus the x of interval 165–169 is $-$ 3.80; and the x of the lowest interval in the distribution, 140–144, is $-$ 28.80.

It will be helpful in calculating deviations from the mean to remember that the mean is *always* subtracted from the individual score or midpoint value. That is, x (deviation) $= X$ (score or midpoint) $- M$(mean). The calculation is algebraic. When the score or midpoint is numerically *larger* than the mean the deviation is positive; when the score or midpoint is numerically *smaller* than the mean the deviation is negative.

Column (4) Table 8, gives the deviation of each class-interval, as represented by its midpoint, from the mean of the distribution. There are more scores on some intervals than on others; hence each midpoint deviation in column (4) must be

"weighted" or multiplied by the number of scores (f) which it represents. This gives the fx column, column (5). The first fx is 26.20; for, since there is only one score on 195–199, we multiply the first x by 1. The next fx is 42.40, since each of the two scores on 190–194 has an x of 21.20. In the same way we obtain the other fx's by multiplying, in each case, the x in column (4) by its corresponding f in column (3). When all of the fx's have been calculated, the column is added without regard to sign, and the resulting sum is divided by N to give the MD. In the present problem the MD equals 502.00/50 or 10.04.

The formula for the MD when measures are grouped into a frequency distribution is as follows:

$$MD = \frac{\Sigma \mid fx \mid}{N} \qquad (8)$$

(mean deviation for scores grouped into a frequency distribution)

The second problem in Table 8 shows the calculation of the MD for 200 cancellation scores grouped into a frequency distribution in class-intervals of four. The mean of this distribution was found to be 119.44 (Table 5). Hence, the x of the topmost interval, 135.5 to 139.5 (midpoint 137.50), from the mean is 18.06. Since the class-interval is constant in size, the next x may be found by subtracting 4 (the interval) from 18.06; and each succeeding x may be found by subtracting 4 from the x just preceding it.

The fx's in column (5) are found, as shown in problem 1, by weighting each x by the f which it represents — by the f opposite it. The sum of the fx column is 1063.88; and, since N is equal to 200, from formula (8) we obtain 5.32 as the MD of the scores in this distribution around their mean of 119.44.

In a symmetrical or normal distribution the MD, when measured off on the scale above and below the mean, marks the limits of the middle 57.5% of the measures. The MD is always slightly larger, therefore, than the Q which marks off the limits of the middle 50%. A large MD means that the

scores of the distribution tend to scatter widely around the central tendency; a small *MD* that they tend to be concentrated within a relatively narrow range.

4. The Standard Deviation or *SD*

The *standard* deviation or *SD* is the measure of variability customarily employed in research. The *SD* differs from the *MD* in several respects. In calculating the *MD* we disregard signs and treat all deviations as positive; in finding the *SD* we avoid this difficulty of signs by squaring the separate deviations. Again, the squared deviations used in computing the *SD* are always taken from the mean of the distribution, and never from the median or mode. The conventional symbol used to denote the *SD* is the Greek letter sigma (σ).

(1) Calculation of *SD* from Ungrouped Data

The standard deviation or σ is the square root of the mean of the squared deviations taken from the arithmetical mean of the distribution. To illustrate the calculation of the *SD* in a simple ungrouped series, let us consider the example given on page 55, to illustrate the calculation of the *MD*, in which the deviations of the five measures, 6, 8, 10, 12, and 14 from their mean of 10 were found to be -4, -2, 0, 2, and 4, respectively. Squaring each of these deviations, we obtain 16, 4, 0, 4, and 16. Summing these five squares and dividing by five, we obtain the mean of the squares, and, extracting the square root, get 2.83, the *SD* of this series. The formula for the *SD* or σ when the series of scores is ungrouped is as follows:

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \tag{9}$$

(*standard deviation calculated from ungrouped data*)

(2) Calculation of *SD* from Grouped Data

Table 8 illustrates the calculation of σ when scores are grouped into a frequency distribution. The process is identical with that used for ungrouped items, except that, in addition to squaring the x of each midpoint from the mean, we weight each

of these squared deviations by the frequency which it represents — that is, by the frequency opposite it. This multiplication gives the fx^2 column. By simple algebra, $x \times fx = fx^2$; and accordingly the easiest way to obtain the entries in column fx^2 is to multiply the corresponding x's and fx's in columns (4) and (5). The first fx^2 entry, for example, is 686.44, the product of 26.20 times 26.20; the second entry is 898.88, the product of 42.40 times 21.20; and so on to the end of the column. All of the fx^2 are necessarily positive since each negative x is matched by a negative fx. The sum of the fx^2 column (7978.00) divided by N (50) gives the mean of the squared deviations as 159.56; and the square root of this result is 12.63, the SD. The formula for σ when data are grouped into a frequency distribution is:

$$\sigma = \sqrt{\frac{\Sigma fx^2}{N}} \cdot \qquad (10)$$

(SD *or* σ *for data grouped into a frequency distribution*)

Problem 2 of Table 8 furnishes another illustration of the calculation of σ from grouped data. In column (6), the fx^2 entries have been obtained, as in the previous problem, by multiplying each x by its corresponding fx. The sum of the fx^2 column is 8927.29; and N is 200. Hence, applying formula (10) we get 6.68 as the SD.

The standard deviation is less affected by *sampling errors* (p. 196) than is the Q or the MD and is a more stable measure of dispersion. In a normal distribution the SD, when measured off above and below the mean, marks the limits of the middle 68.26% (roughly the middle two-thirds) of the distribution. This is approximately true also of the σ in less symmetrical distributions. For example, in the first problem in Table 8 the middle 65% of the scores fall between score 183 (170.80 + 12.63) and score 158 (170.80 − 12.63).* The SD is always larger than the MD which is, in turn, always larger than Q.

* See page 135 for method of calculating the percentage of scores falling between two points in a frequency distribution.

These relationships supply a rough check upon the accuracy of the measures of variability.

II. Calculation of the *SD* by the Short Method

1. Calculation of σ from Grouped Data

On page 41, the Short Method of calculating the mean was outlined. This method consisted essentially in "guessing" or assuming a mean, and later applying to this value a correction to give the actual mean. The Short Method may also be used to advantage in calculating the *SD**. It is a decided time and labor saver in dealing with grouped data; and is well-nigh indispensable in the calculation of σ's in a correlation table (p. 283).

The Short Method of calculating the *SD* is illustrated in Table 9. The computation of the mean is repeated in the table,

TABLE 9

The Calculation of the *SD* by the Short Method.† Data from Table 1. Calculations by the Long Method Given for Comparison

1. Short Method

(1) Scores	(2) Midpoint X	(3) f	(4) x'	(5) fx'		(6) fx'^2
195–199	197	1	5	5		25
190–194	192	2	4	8		32
185–189	187	4	3	12		36
180–184	182	5	2	10		20
175–179	177	8	1	8	(+ 43)	8
170–174	172	10	0			
165–169	167	6	− 1	− 6		6
160–164	162	4	− 2	− 8		16
155–159	157	4	− 3	− 12		36
150–154	152	2	− 4	− 8		32
145–149	147	3	− 5	− 15		75
140–144	142	1	− 6	− 6	(− 55)	36
		$N = 50$		98		322

* The *MD* may also be calculated by the assumed mean or Short Method. The *MD* is so rarely used, however, that the Short Method of calculation (which is neither very short nor very satisfactory) is not given.

† The calculation of the mean is repeated from Table 7.

TABLE 9 (continued)

1. $AM = 172.00 \quad c = -\frac{12}{50} = -.240 \quad ci = -.240 \times 5 = -1.20$
$c^2 = .0576$
$ci = -1.20$
$M = 170.80$

2. $SD = \sqrt{\frac{\Sigma fx'^2}{N} - c^2} \times i \text{ (interval)} = \sqrt{\frac{322}{50} - .0576} \times 5$
$= 12.63$

2. Long Method

(1) Scores	(2) Midpoint X	(3) f	(4) fX	(5) x	(6) fx	(7) fx^2
195–199	197	1	197	26.20	26.20	686.44
190–194	192	2	384	21.20	42.40	898.88
185–189	187	4	748	16.20	64.80	1049.76
180–184	182	5	910	11.20	56.00	627.20
175–179	177	8	1416	6.20	49.60	307.52
170–174	172	10	1720	1.20	12.00	14.40
165–169	167	6	1002	− 3.80	− 22.80	86.64
160–164	162	4	648	− 8.80	− 35.20	309.76
155–159	157	4	628	− 13.80	− 55.20	761.76
150–154	152	2	304	− 18.80	− 37.60	706.88
145–149	147	3	441	− 23.80	− 71.40	1699.32
140–144	142	1	142	− 28.80	− 28.80	829.44
		$N =$ 50	8540		502.00	7978.00

1. $M = \frac{\Sigma fX}{N} = \frac{8540}{50} = 170.80$

2. $SD = \sqrt{\frac{\Sigma fx^2}{N}} = \sqrt{\frac{7978.00}{50}} = 12.63$

as is also the calculation of the mean and SD by the direct or Long Method. This procedure affords a readier comparison of the two techniques.

The formula for computing σ by the Short Method is

$$\sigma = \sqrt{\frac{\Sigma fx'^2}{N} - c^2} \times i \text{ (interval)} \tag{11}$$

(SD *from a frequency distribution when deviations are taken from an assumed mean*)

in which $\Sigma fx'^2$ is the sum of the squared deviations in units of class-interval, taken from the assumed mean, and c^2 is the squared correction in units of class-interval.

The calculation of σ by the Short Method may be followed in detail from Table 9. Deviations are taken from the assumed mean (172.0) in units of class-interval and entered in column (4) as x'. In column (5) each x' is weighted or multiplied by its f to give the fx'; and in column (6) the fx'^2's are found by multiplying each x' in column (4) by the corresponding fx' in column (5). The process is identical with that used in the Long Method except that the x''s are all expressed in units of class-interval. This considerably simplifies the multiplication. The calculation of c has already been described on page 44: c is the algebraic sum of column (5) divided by N. The sum of the fx'^2 column is 322, and c^2 is .0576. Applying formula (11) we get 2.525×5 (interval) or 12.63 as the σ of the distribution. Formula (11) for the calculation of σ by the Short Method holds good no matter what the size of c, the correction in units of class-interval, or where the mean has been assumed.

2. Calculation of σ from the Original Measures or Scores

It will often save time and labor to apply the Short Method for computing σ directly to the ungrouped scores. The method is illustrated in Table 10. Note that the ten scores are ungrouped, and that it is not necessary even to arrange them in order of size. The assumed mean is taken at zero, and each score becomes at once a deviation (x') from this AM, that is, each score (X) is unchanged. The correction, c, is the difference between the actual mean (M) and the assumed mean (0), i.e., $c = M - 0$; hence c is simply M itself. The mean is calculated, as before, by summing the scores and dividing by N (see p. 32). To find σ, we square the x''s (or the X's which are the scores), sum them to get $\Sigma(x')^2$ or ΣX^2, divide by N, and subtract M^2, the correction squared. The square root of the result gives σ. A convenient formula is

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - M^2} \tag{12}$$

or replacing the M^2 by $\left(\frac{\Sigma X}{N}\right)^2$,

$$\sigma = \frac{\sqrt{N\Sigma X^2 - (\Sigma X)^2}}{N} \tag{13}$$

(*σ calculated from original scores by the Short Method*)

This method of calculating σ is especially useful when there are relatively few scores, say fifty or less, and when the scores are expressed in not more than two digits,* so that the squares do not become unwieldy. A calculating machine and a table of squares will greatly facilitate computation. Simply sum the scores as they stand and divide by N to get M. Then enter the squares of the scores in the machine in order, sum, and substitute the result in formula (12) or formula (13).

TABLE 10

To Illustrate the Calculation of the *SD* from Original Scores When the Assumed Mean Is Taken at Zero, and Data Are Ungrouped

Scores (X)	x' (or X)	$(x')^2$ or (X^2)
18	18	324
25	25	625
21	21	441
19	19	361
27	27	729
31	31	961
22	22	484
25	25	625
28	28	784
20	20	400
236	236	5734

$$\begin{aligned}
AM &= 0 \\
M &= \tfrac{236}{10} = 23.6 \qquad\qquad N = 10 \\
c &= 23.6 - 0 \\
&= 23.6 \\
c^2 &= 557.96 \\
\sigma &= \sqrt{\tfrac{5734}{10} - (23.6)^2} \times 1 \text{ (interval)} \\
&= \sqrt{16.44} \\
&= 4.06
\end{aligned}$$

* For the application of this method to the calculation of coefficients of correlation, and a scheme for reducing the size of the original scores so as to eliminate the need for handling large numbers, see page 293.

3. Effect upon σ of (a) Adding a Constant to Each Score, or (b) Multiplying Each Score by the Same Number

(a) If each score in a frequency distribution is increased by some set amount, say 5, the σ is unchanged. The table below provides a simple illustration. The mean of the original scores is 7 and σ is 1.41. When each score is increased by 5, the mean is 12 $(7 + 5)$, but σ is still 1.41. Adding a constant (e.g. 5, 10, 15) to each score simply moves the whole distribution up the scale 5, 10, or 15 points. The mean is increased by the amount of the constant added, but the variability (σ) is not affected. If a constant is subtracted from each score, the distribution is moved down the scale by that amount; the mean is decreased by the amount of the constant, and σ, again, is unchanged.

Original scores (X)	x	x^2	Original scores $X + 5$	x	x^2
9	2	4	14	2	4
8	1	1	13	1	1
7	0	0	12	0	0
6	− 1	1	11	− 1	1
5	− 2	4	10	− 2	4
5)35		10	5)60		10
$M = 7$			$M = 12$		

$\sigma = \sqrt{\frac{10}{5}} = 1.41$ $\qquad$ $\sigma = \sqrt{\frac{10}{5}} = 1.41$

(b) What happens to σ when each score is multiplied by a constant is shown in the table below:

Original scores (X)	Original scores $X \times 10$	x	x^2
9	90	20	400
8	80	10	100
7	70	0	0
6	60	− 10	100
5	50	− 20	400
35	5)350		1000
$M = 7$	$M = 70$		
$\sigma = 1.41$			

$$\sigma = \sqrt{\frac{1000}{5}} = \sqrt{200} = 14.14$$

Each score in the list of five, shown above, has been multiplied by 10. It is evident that the net effect of this operation has been to multiply the mean *and* the σ by 10.

III. The Coefficient of Variation, V

It is often desirable to compare the variability of a given group upon two or more *different* tests; or to compare the variabilities of two or more groups upon the *same* test. We may wish, for example, to know whether eight-year-old girls are more variable in height than in weight; or whether ten-year-old boys are more variable than ten-year-old girls in vocabulary or in memory span. The Q, MD, and SD are not suitable, ordinarily, for such comparisons. These measures give the *absolute* spread or dispersion of test scores around their means in terms of the units of the test. But owing to differences in measuring units, we cannot compare the variability in height and the variability in weight of a given group directly; nor can we compare the relative variability in height of two groups, say boys and girls, unless the means of the two distributions are at least approximately equal. To enable us to tell whether one group is more variable than another, we need a measure which takes account *both* of the central tendency *and* of the variability of the group, and which is independent of the units in which ability is expressed. One such measure is the ratio σ/M, called the *coefficient of variation*, or V. The formula for V is

$$V = \frac{100 \times \sigma}{M} \tag{14}$$

(*the coefficient of variation or coefficient of relative variability*) *

The following illustrations will make the use of the formula clear. Consider, first, the case where abilities are measured in different units. A group of seven-year-old boys has a mean height of 45 inches with a σ of 2.5 inches; and a mean weight

* The multiplier 100 is introduced for the purpose of avoiding small fractional results.

of 50 pounds with a σ of 6.0 pounds. In which trait is the group more variable, height or weight? Since we cannot compare inches and pounds directly, it is impossible to answer this question by reference to the *SD*'s of the height and weight distributions. But we can compare the relative variability of the two distributions in terms of their coefficients of variation. Thus,

$$V_{ht} = \frac{100 \times 2.5}{45} = 5.6 \qquad \text{by formula (14)}$$

and

$$V_{wt} = \frac{100 \times 6.0}{50} = 12 \qquad \text{by formula (14)}$$

from which it appears that these boys are 5.6/12 or 47% as variable in height as in weight.

Now let us consider the case where variability is measured in the same units, but around different points on the scale. At the end of five minutes, a group of fifty children had worked an average of 20.50 examples correctly, the σ being 5.24. At the end of ten minutes, the same group had worked an average of 34.80 examples correctly, the σ being 9.62. If we compared the σ's of the two distributions directly, we should probably be inclined to conclude that the group was nearly twice as variable at the end of the ten-minute period as it was at the end of the five-minute period, since the σ has increased from 5.24 to 9.62. This conclusion is correct as far as the *absolute* spread or variability within the group is concerned. But to compare the *relative* dispersion of the group in the two periods, we must take account of the fact that, with the increase in σ, the means have also increased from 20.50 to 34.80. The coefficients of variation give the following results:

$$\text{For the five-minute period: } V = \frac{100 \times 5.24}{20.50} = 25.6$$

$$\text{For the ten-minute period: } V = \frac{100 \times 9.62}{34.80} = 27.6$$

Thus, instead of being about 50% as variable in the five-minute period as in the ten, the group is 25.6/27.6 or 93% as variable,

when the mean score is considered as well as the absolute variability.

Objection has been raised* to the use of V in comparing the relative variability of test scores because the "true" zero point of ability in mental and educational tests is unknown. This objection does not apply, of course, to physical and physiological measures since these have true zeros. How the lack of knowledge of the true zero in a mental test may affect V can be shown most readily, perhaps, by an example. Suppose that we have given a vocabulary test to a group of children, and have obtained a mean of 25 and a σ of 5. V will equal 20. Now suppose that we add 30 very easy items, say, to our vocabulary test. It is highly probable that every child will know all of the added words, and hence the mean score as well as every subject's score will be increased by 30. The absolute variability of the group (the σ) will, however, remain unchanged, as each subject occupies exactly the same relative position as before. An increase in the mean (from 25 to 55) without a corresponding increase in σ changes V from 20 to 9; and, since we could add 40 or 400 items as easily as 30, V appears to be a very unstable measure.

While theoretically correct, criticism of V because of the arbitrary nature of the zero point in mental and educational tests is not so generally destructive as it seems. Makers of standard psychological tests have been careful to begin their tests with items which, by experimental tryout, have been found to have minimal difficulty for the group for whom the test is designed. While admittedly arbitrary, such "zero" points are at least located at extremely low levels of difficulty in the ability measured by the test; hence it would be foolish to include additional easy items at the low end of the scale. The mean tells us how far the group has progressed, on the average, from the arbitrary zero point of the test. V shows,

* Franzen, R., "Statistical Issues," *Journal of Educational Psychology,* 15 (1924), 367–382.

Thurstone, L. L., "The Absolute Zero in Intelligence Measurement," *Psychological Review,* 35 (1928), 175–197.

essentially, what percentage the variability is of this distance. Like M, V has a definite meaning for the test as it stands. If the range of difficulty in the test is altered, or the units changed, not only V, but M, is changed. V, therefore, is in a sense no more arbitrary than M, and the objections raised against this measure can be directed with equal force against M.

V is most useful, perhaps, in comparing the variability of a group upon the *same* test administered under different conditions, as, for example, when a group of students works at a task with and without distraction. The zero point here, at least, remains substantially constant. V may also be used to compare two or more groups on the *same* test, as when ten-year-old boys and ten-year-old girls are compared in tests of logical memory or picture completion. In both of these cases it is probably justifiable to assume that the "true" zero point of ability is sensibly the same for the groups compared.

It is, perhaps, most difficult to interpret V when the variability of a group upon *different* mental tests is a matter of interest. If we compare a group of girls for variability in paragraph reading and in arithmetic computaton, it should be made plain that the V's refer *only* to the specific scales upon which performance has been measured. Other tests of reading and arithmetic may — and probably will — give different results because of difference in test units, range of difficulty covered by the test, and position of arbitrary zero points. But if one restricts his use of V to the particular measures which he has employed, this coefficient will furnish useful information.

IV. The Short Method Applied to Discrete Series

We have defined a truly discrete series on page 2 as one in which there are real gaps. This means that in a discrete series each measure, instead of representing an interval on a scale as in a continuous series, is a separate and distinct value. There is, for example, a real gap between one man and two men; or between one dollar and two dollars, provided the unit of measurement in the latter case is one dollar.

Table 11 illustrates the method of calculating the measures of central tendency and variability for discrete measures tabulated into a frequency distribution. The data consist of the records of the number of children in forty-four families in a rural community. In the first column of the table is given the number of children in the family; in the second column — under f — the number of families of a given size. We find, for instance, one family of ten children; three of nine; four of eight, etc. Since the measures — here, the children — are discrete, each measure must be taken at face value, and there are, in consequence, no midpoint values for the different steps.

TABLE 11

To Illustrate the Calculation of the Mean, the Median, Q and SD When Measures Are Discrete

(Note that the f column gives the number of families containing the children listed in the first column)

Number of Children	Families f	x'	fx'	fx'^2
10	1	5	5	25
9	3	4	12	48
8	4	3	12	36
7	3	2	6	12
6	5	1	5 (+ 40)	5
5	8	0		
4	7	− 1	− 7	7
3	4	− 2	− 8	16
2	4	− 3	− 12	36
1	2	− 4	− 8	32
0	3	− 5	− 15 (− 50)	75
	$N = 44$		90	292

$AM = 5.00$ $\quad c = \frac{-10}{44} = -.23$ $\quad c^2 = .053$

$ci = -.23$

$M = 4.77$ $\quad N/2 = 22$; and, since the 22nd measure falls on 5, the $Mdn = 5$

$Mdn = 5$ $\quad N/4 = 11$; and, since the 11th measure falls on 3, $Q_1 = 3$

$Mode = 5$ $\quad 3N/4 = 33$; and, since the 33rd measure falls between 6 and 7, $Q_3 = 6.5$

$$Q = \frac{6.5 - 3}{2} = 1.75$$

$$SD = \sqrt{\tfrac{292}{44} - .053} \times 1 \text{ (interval)} = 2.57$$

The mean is guessed at 5, and x''s are taken directly from this point. The fx' and the fx'^2 columns are calculated exactly as shown in Table 9 for a continuous series — the first column is obtained by multiplying the corresponding f and x' values, and the second by multiplying corresponding x' and fx' values. Since the class-interval is 1, the correction c equals ci directly.

If we apply the correction $-.23$ to 5.00 (the "guessed" mean), 4.77, the mean of the distribution, is obtained. This result, while mathematically correct, is rather difficult to interpret in a practical way, as it is obviously impossible for a family to have four and a fraction of children. Is the median a more meaningful measure? One-half of the measures is 22, and counting in from the small end of the series we find that the twenty-second score falls on interval 5. Fractional values are, of course, really meaningless in a discrete series; and hence we simply take 5 as being roughly the median of the distribution without any interpolation. The median family, accordingly (and the modal family as well), may be said to contain five children, and this result on the face of it is of greater utility than the statement that the average number of children in a family is 4.77.

In computing measures of variability in a discrete series, the Q is the only one which offers difficulties. In the present illustration, one-fourth ($N/4$) of the measures is 11, and, counting in from the low end of the series eleven scores, we put Q_1 on 3 (as in the case of the median, no interpolation is made). If we check this value of Q_1 by counting in thirty-three scores from the high end of the distribution, we again obtain 3 as the value of Q_1. Three-fourths ($3N/4$) of the measures is 33; and, counting in thirty-three scores from the low end, we complete — or count through — the frequency on 6. If eleven scores are counted off from the other direction, we complete — or count through — the frequency on 7. This puts Q_3 at either 6 or 7, and the best way out of the difficulty is to take Q_3 as roughly equal to 6.5, i.e., midway between 6 and 7. Taking Q_1 equal to 3, and Q_3 equal to 6.5, Q is $\frac{6.5-3}{2}$ or 1.75.

The σ in a discrete series is found from formula (11) in exactly the same way as in a continuous series. In Table 11, the σ is $\sqrt{\frac{292}{44} - .053} \times 1$ (the class-interval) or 2.57.

V. When to Use the Various Measures of Variability

1. *Use the range*
   (1) When the data are too scant or too scattered to justify the calculation of any other measure of variability.
   (2) When a knowledge of the total spread of scores is all that is wanted.

2. *Use the Q*
   (1) For a quick, inspectional measure of variability.
   (2) When there are scattered or extreme measures.
   (3) When the degree of concentration around the median is sought.

3. *Use the MD*
   (1) When it is desired to weight all deviations according to their size.
   (2) When extreme deviations should influence the measure of variability, but not influence it unduly.

4. *Use the SD*
   (1) When the measure having the highest degree of reliability is sought (p. 196).
   (2) When it is desired that extreme deviations have a proportionally greater influence upon the measure of variability.
   (3) When coefficients of correlation or measures of reliability are subsequently to be computed (p. 282).

PROBLEMS

1. Calculate the Q and σ for each of the four frequency distributions given on page 46 under problem 1, Chapter II.

2. Calculate the σ of the twenty-five ungrouped scores given on page 28, problem 5(a), taking the AM at zero. Compare your result with the σ's calculated from the frequency distributions of the same scores which you tabulated in class-intervals of three and five units.

3. For the following list of test scores,

52, 50, 56, 68, 65, 62, 57, 70

(*a*) Find the M and σ by method on page 60.
(*b*) Add 6 to each score and recalculate M and σ.
(*c*) Subtract 50 from each score, and calculate M and σ.
(*d*) Multiply each score by 5 and compute M and σ.

4. Calculate coefficients of variation for the following traits:

Trait	Unit of measurement	Group	M	σ
Length of Head	mms.	802 males	190.52	5.90
Body Weight	pounds	868,445 males	141.54	17.82
Tapping Speed	M of 5 trials 30″ each	68 adults, male and female	196.91	26.83
Memory Span	No. repeated correctly	263 males	6.60	1.13
General Intelligence (Otis Group Intell. Scale)	Points scored	1101 adults	153.3	23.6

Rank these traits in order for relative variability. Judged by their V's which trait is the most variable? which the least variable? which traits have true zeros?

5. (*a*) Why is the Q the best measure of variability when there are scattered or extreme scores?
(*b*) Why does the σ weight extreme deviations more than does the MD?

Answers

1. (1) $Q = 3.38$, $\sigma = 4.99$ (2) $Q = 8.13$, $\sigma = 11.33$
(3) $Q = 4.50$, $\sigma = 7.23$ (4) $Q = 16.41$, $\sigma = 24.13$

2. σ of ungrouped scores = 6.72
σ of scores grouped in 3-unit intervals = 6.71
σ of scores grouped in 5-unit intervals = 6.78

3. (*a*) $M = 60$ (*b*) $M = 66$ (*c*) $M = 10$ (*d*) $M = 300$
$\sigma = 6.91$ $\sigma = 6.91$ $\sigma = 6.91$ $\sigma = 34.55$

4. *V*'s in order are 3.10; 12.59; 13.63; 17.12; 15.39. Ranked for relative variability from most to least: Memory Span; General Intelligence; Tapping Speed; Weight; Head Length. Last two traits have true zeros.

CHAPTER IV

CUMULATIVE DISTRIBUTIONS, GRAPHIC METHODS, AND PERCENTILES

IN Chapter I, we learned how to represent the frequency distribution by means of the polygon and the histogram. In the present chapter, other descriptive methods will be considered — the *cumulative frequency graph*, the *cumulative percentage curve* or *ogive*, and certain simple graphical devices. Also, methods will be given for calculating percentiles and percentile ranks from frequency distributions and directly from graphs.

I. THE CUMULATIVE FREQUENCY GRAPH

1. Construction of the Cumulative Frequency Graph

The cumulative frequency graph is another way of representing a frequency distribution by means of a diagram. Before we can plot a cumulative frequency graph, the scores of the distribution must be added serially or cumulated, as shown in Table 12, for the two distributions taken from Table 5, page 35. These two sets of scores have already been used to illustrate the frequency polygon and histogram in Figures 2, 4, and 5. The first two columns for each of the distributions in Table 12 repeat Table 5, page 35, exactly; but in the third column (Cum. f) scores have been "accumulated" progressively from the bottom of the distribution upward. To illustrate, in the distribution of Army Alpha scores the first "cumulative frequency" is 1; $1 + 3$, from the low end of the distribution, gives 4 as the next entry; $4 + 2 = 6$; $6 + 4 = 10$, etc. The last cumulative frequency is, of course, equal to 50 or N, the total frequency.

The two cumulative frequency graphs which represent the

TABLE 12

CUMULATIVE FREQUENCIES FOR THE TWO DISTRIBUTIONS GIVEN IN TABLE 5, P. 35

Army Alpha Scores	*f*	Cum. *f*	Cancellation Scores	*f*	Cum. *f*
195–199	1	50	135.5 to 139.5	3	200
190–194	2	49	131.5 to 135.5	5	197
185–189	4	47	127.5 to 131.5	16	192
180–184	5	43	123.5 to 127.5	23	176
175–179	8	38	119.5 to 123.5	52	153
170–174	10	30	115.5 to 119.5	49	101
165–169	6	20	111.5 to 115.5	27	52
160–164	4	14	107.5 to 111.5	18	25
155–159	4	10	103.5 to 107.5	7	7
150–154	2	6		$N = 200$	
145–149	3	4			
140–144	1	1			
	$N = 50$				

distributions of Table 12 are shown in Figures 8 and 9. Consider first the graph of the fifty Army Alpha scores in Figure 8. The class-intervals of the distribution have been laid off along the *X-axis.* There are twelve intervals, and by the "75% rule" given on page 13 there should be about nine unit distances (each equal to one class-interval) laid off on the *Y-axis.* Since the largest cumulative frequency is 50, each of these *Y*-units should represent 50/9 or 6 scores (approximately). Instead of dividing up the total *Y*-distance into nine units each representing six scores, however, we have, for convenience in plotting, divided the total *Y*-distance into ten units of five scores each. This does not change significantly the 3:4 relationship of height to width in the figure.

When plotting the frequency polygon the frequency on each interval is taken at the *midpoint* of the class-interval. But in constructing a cumulative frequency curve each cumulative frequency is plotted at the *upper limit* of the interval upon which it falls. This is because we are adding progressively from bottom up and hence each cumulative frequency carries through to the upper limit of the interval. The first point on the curve is one *Y*-unit (the cumulative frequency on 140–144) just above 144.5; the second point is four *Y*-units just

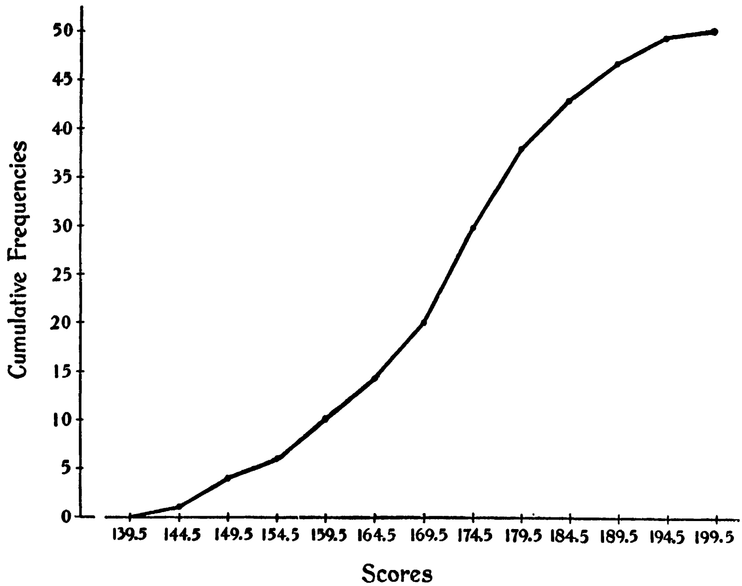


FIG. 8. Cumulative Frequency Graph. (Data from Table 12, p. 75.)

above 149.5; the third, six Y-units just above 154.5, and so on to the last point which is fifty Y-units above 199.5. The plotted points are joined to give the S-shaped cumulative frequency graph. In order to have the curve begin on the X-*axis* it is started at 139.5 (upper limit of 134.5 to 139.5), the cumulative frequency of which is 0.

The cumulative frequency curve in Figure 9 has been plotted from the second distribution in Table 12 by the method just described. The curve begins at 103.5, the lower limit of the first class-interval,* and ends at 139.5, the upper limit of the last interval; and cumulative frequencies, 7, 25, 52, etc., are all plotted at the *upper limits* of their respective class-intervals. The height of this graph was determined by the "75% rule" as in the case of the curve in Figure 8. There are nine class-intervals laid off on the X-*axis;* hence, since 75% of 9 is 7

* Or the upper limit of the interval just below, i.e., 99.5 to 103.5.

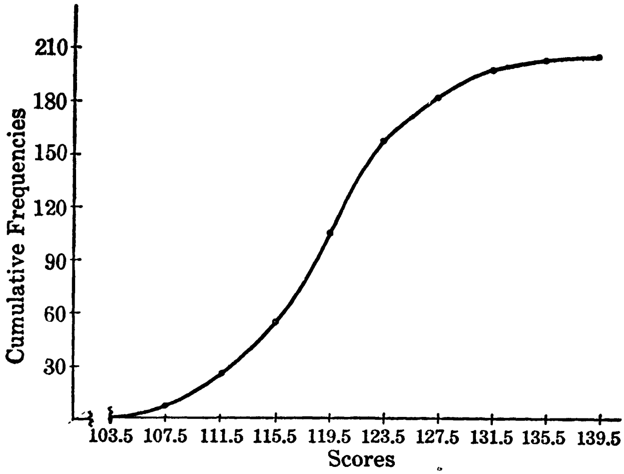


FIG. 9. Cumulative Frequency Graph. (Data from Table 12, p. 75.)

(approximately), the height of the figure should be about seven class-interval units. To determine the score value of each Y-unit divide 200 (the largest cumulative frequency) by 7 to give 30 (approximately). Each of the seven Y-units has been taken to represent 30 scores.

II. PERCENTILES AND PERCENTILE RANKS

1. Calculation of Percentiles in a Frequency Distribution

We have learned (p. 36) that the median is that point in a frequency distribution below which lie 50% of the measures or scores; and that Q_1 and Q_3 mark points in the distribution below which lie, respectively, 25% and 75% of the measures or scores. In exactly the same way in which the median and quartiles are found, we may compute points below which lie 10%, 43%, 85%, or any "precent" of the scores. These points are called *percentiles*, and are designated, in general, by the symbol P_p, the p referring to the percentage of cases *below* the given value. P_{10}, for example, is the point below which lie 10% of the scores; P_{78}, the point below which lie 78% of the scores. It

is evident that the median, expressed as a percentile, is P_{50}; also Q_1 is P_{25}, and Q_3 is P_{75}.

The method of calculating percentiles is essentially the same as that employed in finding the median. The formula is

$$P_p = l + \left(\frac{pN - F}{f_p}\right) \times i \text{ (interval)} \tag{15}$$

(percentiles in a frequency distribution, counting from below up)

where

p = percentage of the distribution wanted, e.g., 10%, 33%, etc.

l = lower limit of the class-interval upon which P_p lies

pN = part of N to be counted off in order to reach P_p

F = sum of all scores upon intervals below l

f_p = number of scores within the interval upon which P_p falls

i = length of the class-interval

In Table 13, the percentile points, P_{10} to P_{90}, have been computed by formula (15) for the distribution of scores made by the fifty college students upon Army Alpha, shown in Table 1, page 6. The details of calculation are given in Table 13. We may illustrate the method with P_{70}. Here, $pN = 35$ (70% of 50 = 35), and from the Cum. f we find that 30 scores take us through 170–174 up to 174.5, the *lower* limit of the interval next above. Hence, P_{70} falls upon 175–179, and, substituting $pN = 35$, $F = 30$, $f_p = 8$ (frequency upon 175–179), and $i = 5$ (class-interval) in formula (15), we find that $P_{70} = 177.6$ (for detailed calculation, see Table 13). This result means that 70% of the fifty students scored *below* 177.6 in the distribution of Army Alpha scores. The other percentile values are found in exactly the same way as P_{70}. The reader should verify the calculations of the P_p in Table 13 in order to become thoroughly familiar with the method.

It should be noted that P_0, which marks the lower limit of the first interval (namely, 139.5) lies at the beginning of the distribution. P_{100} marks the *upper limit* of the last interval,

TABLE 13

CALCULATION OF CERTAIN PERCENTILES IN A FREQUENCY DISTRIBUTION

(Data are fifty Army Alpha scores, see Table 1, p. 6)

Scores	f	Cum. f	Percentiles
195–199	1	50	$P_{100} = 199.5$
190–194	2	49	
185–189	4	47	$P_{90} = 187.0$
180–184	5	43	$P_{80} = 181.5$
175–179	8	38	$P_{70} = 177.6$
170–174	10	30	$P_{60} = 174.5$
165–169	6	20	$P_{50} = 172.0$
160–164	4	14	$P_{40} = 169.5$
155–159	4	10	$P_{30} = 165.3$
150–154	2	6	$P_{20} = 159.5$
145–149	3	4	$P_{10} = 152.0$
140–144	1	1	
	$N = 50$		$P_0 = 139.5$

CALCULATION OF PERCENTILES (DECILE POINTS)

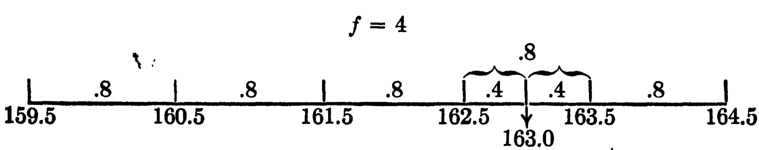
10% of 50 = 5	$149.5 + \left(\frac{5-4}{2}\right) \times 5 = 152.0$
20% of 50 = 10	$159.5 + \left(\frac{10-10}{4}\right) \times 5 = 159.5$
30% of 50 = 15	$164.5 + \left(\frac{15-14}{6}\right) \times 5 = 165.3$
40% of 50 = 20	$169.5 + \left(\frac{20-20}{10}\right) \times 5 = 169.5$
50% of 50 = 25	$169.5 + \left(\frac{25-20}{10}\right) \times 5 = 172.0$ (*Mdn*)
60% of 50 = 30	$174.5 + \left(\frac{30-30}{8}\right) \times 5 = 174.5$
70% of 50 = 35	$174.5 + \left(\frac{35-30}{8}\right) \times 5 = 177.6$
80% of 50 = 40	$179.5 + \left(\frac{40-38}{5}\right) \times 5 = 181.5$
90% of 50 = 45	$184.5 + \left(\frac{45-43}{4}\right) \times 5 = 187.0$

and lies at the end of the distribution. These two percentiles represent *limiting points*. Their principal value is to indicate the boundaries of the percentile scale.

2. Calculation of Percentile Ranks in a Frequency Distribution

We have seen in the last section how percentiles, e.g., P_{15} or P_{62}, may be calculated directly from a frequency distribution. To repeat what has been said above, percentiles are *points* in a continuous distribution below which lie given percentages of N. We shall now consider the problem of finding an individual's *percentile rank* (PR); or the position on a scale of 100 to which the subject's score entitles him. The distinction between *percentile* and *percentile rank* will be clear if the reader remembers that in calculating percentiles he *starts* with a certain percent of N, say 15% or 62%. He then counts into the distribution the given percent and the point reached is the required percentile, e.g., P_{15} or P_{62}. The procedure followed in computing percentile ranks is the reverse of this process. Here we begin with an individual *score*, and determine the percentage of scores which lies below it. If this percentage is 62, say, the score has a percentile rank or PR of 62 on a scale of 100.

We may illustrate with Table 13. What is the PR of a man who scores 163? Score 163 falls on interval 160–164. There are ten scores up to 159.5, lower limit of this interval (see column Cum. f), and four scores spread over this interval. Dividing 4 by 5 (interval length) gives us .8 score *per unit of interval.* The score of 163, which we are seeking, is 3.5 score units from 159.5, lower limit of the interval within which the score of 163 lies. Multiplying 3.5 by .8 we get 2.8 as the score-distance of 163 from 159.5; and adding 2.8 to 10 (number of scores below 159.5) we get 12.8 as the part of N lying *below* 163. Dividing 12.8 by 50 gives us 25.6% as that proportion of N below 163; hence the percentile rank of score 163 is 26. The diagram below will clarify the calculation:



Ten scores lie below 159.5. Prorating the four scores on 160–164 over the interval of 5, we have .8 score per unit of interval. Score 163 is just .8 + .8 + .8 + .4 or 2.8 scores from 159.5; or score 163 lies 12.8 scores or 25.6% (12.8/50) into the distribution.

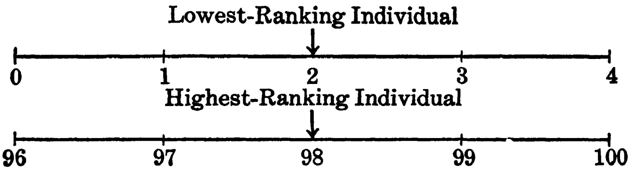
The *PR* of any score may be found in the same way. For example, the percentile rank of 181 is 79 (verify it). The reader should note that a score of 163 is taken as 163.0, midpoint of the score-interval 162.5 to 163.5. This means simply that the midpoint is assumed to be the most representative value in a score-interval. The percentile ranks for several scores may be read directly from Table 13. For instance, 152 has a *PR* of 10; 172 (median) a *PR* of 50, and 187 a *PR* of 90. If we take the percentile-points as representing *approximately* the score-intervals upon which they lie, the *PR* of 160 (upon which 159.5 lies) is approximately 20 (see Table 13); the *PR* of 165 (upon which 165.3 lies) is approximately 30; the *PR* of 170 is approximately 40; of 175, 60; of 178, 70; of 182, 80. These *PR*'s are not strictly accurate, to be sure, but the error is slight.

3. Calculation of Percentile Ranks When Individuals or Objects Are in Order of Merit

Percentile ranks are often used in experimental psychology when we are dealing with attributes for which individuals or objects may be arranged in order of merit, but in which they cannot be measured directly. Children, for instance, may be arranged in order of merit for inventiveness or for social adjustment, pictures and musical selections may be ranked for aesthetic qualities, compositions and handwriting specimens for excellence. When translated over into *PR*'s these ranks may be treated as scores (p. 174).

We may illustrate the use of *PR*'s in such situations for the simple case of twenty-five officer candidates ranked 1, 2, 3, ... 25 in order of merit for "leadership qualities." Here the highest-ranking man has a percentile rank of 98; and the lowest-ranking man a percentile rank of 2. How these values are calculated may be shown in the following way: On a scale run-

ning from 0 to 100, each of twenty-five individuals occupies four divisions (100/25 or 4%) of the scale. Hence, we assign to the poorest individual the *midpoint* of the first four divisions on the scale (0–4) or 2; to the next poorest, the midpoint of the next four divisions (4–8) or 6; and to the best person, the midpoint of the four highest divisions (96–100) or 98. Diagrams illustrating the method of assigning percentile ranks to the best and poorest persons in a group of twenty-five will make the procedure clearer:



If 100 people are arranged in order of merit, what is the percentile rank of the lowest-ranking person? The answer is clear. Since there are just 100 subjects, each occupies one division (100/100 or 1%) on the percentage scale. Hence, the rank of the poorest subject is .5 (midpoint of the interval 0–1) and of the best subject 99.5 (midpoint of the interval 99–100). These *PR*'s and those of the last example may be readily found by means of the following formula* which converts orders of merit into equivalent percentile ranks:

$$PR = 100 - \frac{(100R - 50)}{N} \tag{16}$$

(percentile ranks for individuals ranked in order of merit)

The R in the formula is the rank position of the individual counting #1 as the highest rank in the group. Thus, the individual who ranks highest, i.e., #1, in a group of twenty-five has a PR of $100 - \frac{(100 \times 1 - 50)}{25}$ or 98; and the individual who

* For a table giving percentile ranks for scores ranked in order of merit, and ranging from 11 to 100 in number, see Buros, F. C., and Buros, O. K.. *Expressing Educational Measures as Percentile Ranks*, Test Method Helps, #3, (Yonkers, N.Y.: World Book Co., 1930). In this table a rank of 1 is taken to be the highest, of 2 the next highest, etc.

ranks fifth (i.e., five from the top, twenty from the bottom) has a PR of $100 - \frac{(100 \times 5 - 50)}{25}$ or 82. The person who ranks fiftieth in a group of 100 has a PR of $100 - \frac{(100 \times 50 - 50)}{100}$ or 50.5, the middle of interval 50–51 on the percent scale. Since a person's percentile rank is always the midpoint of an interval on the scale which runs from 0 to 100, it is evident that no one can have a percentile rank of 0 or of 100. These two points constitute the limits of the percentile scale.

III. The Cumulative Percentage Curve or Ogive

1. Construction of the Ogive

The cumulative percentage curve or ogive differs from the cumulative frequency graph in that frequencies are expressed as cumulative *percents* of N on the Y*-axis* instead of as cumulative scores. Table 14 shows how cumulative frequencies are expressed as percentages of N. The distribution consists of

TABLE 14

Calculation of Cumulative Percentages to Upper Limits of Class-Intervals in a Frequency Distribution

(The data represent scores on a reading test achieved by 125 seventh-grade children)

(1) Scores	(2) f	(3) Cum. f	(4) Cum. Percent f
74.5 to 79.5	1	125	100.0
69.5 to 74.5	3	124	99.2
64.5 to 69.5	6	121	96.8
59.5 to 64.5	12	115	92.0
54.5 to 59.5	20	103	82.4
49.5 to 54.5	36	83	66.4
44.5 to 49.5	20	47	37.6
39.5 to 44.5	15	27	21.6
34.5 to 39.5	6	12	9.6
29.5 to 34.5	4	6	4.8
24.5 to 29.5	2	2	1.6
	$N = 125$		

$$\text{Rate} = \frac{1}{N} = \frac{1}{125} = .008$$

scores made on a reading test by 125 seventh-grade pupils. In columns (1) and (2) class-intervals and frequencies are listed; and in column (3) the f's have been cumulated from the low end of the distribution upward as described before on page 74. These Cum. f's are expressed as percentages of N (125) in column (4). The conversion of Cum. f's into cumulative percents can be carried out by dividing each cumulative f by N; e.g., $2 \div 125 = .016$, $6 \div 125 = .048$, and so on. A better method — especially when a calculating machine is available — is to determine first the reciprocal, $1/N$, called the *Rate*, and multiply each cumulative f in order by this fraction. As shown in Table 14, the *Rate* is $1/125$ or .008. Hence, multiplying 2 by .008, we get .016 or 1.6%; $6 \times .008 = .048$ or 4.8%; $12 \times .008 = .096$ or 9.6%, etc.

The curve in Figure 10 represents an ogive plotted from the data in column (4), Table 14. Class-intervals have been laid off on the *X-axis*, and a scale consisting of ten equal distances, each representing 10% of the distribution, has been marked off on the *Y-axis*. The first point on the ogive is placed 1.6 Y-units just above 29.5; the second point is 4.8 Y-units just above 34.5, etc. The last point is 100 Y-units above 79.5, upper limit of the highest class-interval.

2. Computing Percentiles and Percentile Ranks from (*a*) the Cumulative Percentage Distribution and from (*b*) the Ogive

(*a*) Percentiles may be readily determined by direct interpolation in column (4), Table 14. We may illustrate by calculating the 71st percentile. Direct interpolation between the percentages in column (4) gives the following:

	66.4% of the distribution up to 54.5	
71.0% (given) ------→	-----------------------------------	→ 55.9
	82.4% of the distribution up to 59.5	
	16.0% 5.0	

The 71st percentile lies 4.6% above 66.4%. By simple proportion, $\frac{4.6}{16.0} = \frac{x}{5}$ or $x = \frac{4.6}{16.0} \times 5 = 1.4$ (x is the distance of the

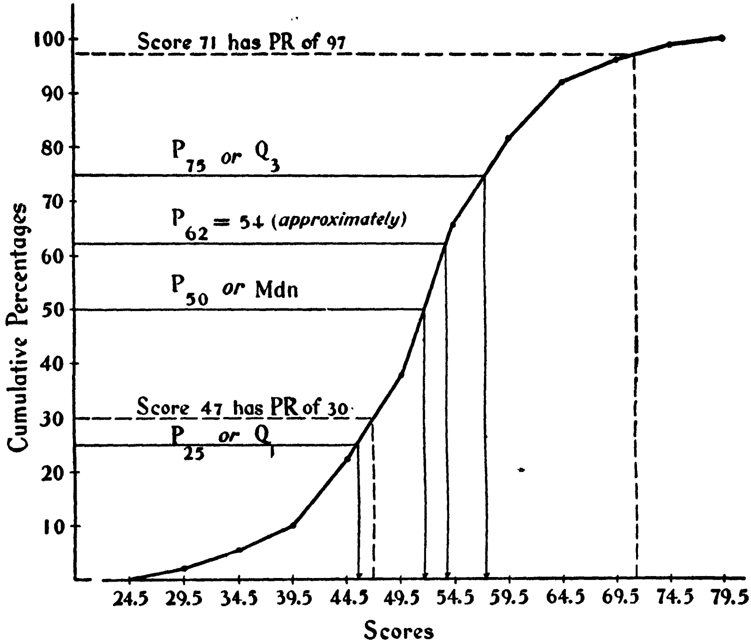


FIG. 10. Cumulative Percentage Curve or Ogive Plotted from the Data of Table 14.

71st percentile from 54.5). The 71st percentile, therefore, is 54.5 + 1.4, or 55.9.

Certain percentiles can be read directly from column (4). We know, for instance, that the 5th percentile is approximately 34.5; that the 22nd percentile is approximately 44.5; that the 38th percentile is approximately 49.5; and that the 92nd percentile is exactly 64.5. Another way of expressing the same facts is to say that 21.6% of the seventh graders scored below 44.5, that 92% scored below 64.5, etc.

Percentile ranks may also be determined from Table 14 by interpolation. Suppose, for example, we wish to calculate the *PR* of score 43. From column (4) we find that 9.6% of the scores are *below* 39.5. Score 43 is 3.5 (43.0 − 39.5) from this

point. There are five score-units on the interval 39.5 to 44.5 which correspond to 12.0% (21.6 − 9.6) of the distribution; hence, 3.5/5 × 12.0 or 8.4 is the percentage distance of score 43 from 39.5. Since 9.6% (up to 39.5) + 8.4% (from 39.5 to 43.0) comprise 18% of the distribution, this percentage of N lies *below* score 43. Hence, the PR of 43 is 18. See detailed calculation below.

$$\begin{array}{lcr}
 & 9.6\% \text{ of distribution up to } 39.5 & \\
18.0\% \longleftarrow\text{-----------} & \longleftarrow\text{-----} & \text{score } 43.0 \\
 & \underline{21.6\%} \text{ of distribution up to } \underline{44.5} & \text{(given)} \\
 & 12.0\% \qquad\qquad\qquad\qquad 5.0 &
\end{array}$$

Score 43.0 is 3.5/5 × 12.0% or 8.4% from 39.5; hence score 43.0 is 9.6% + 8.4% or 18.0% into the distribution.

It should be noted that the cumulative percents in column (4) give the PR's of the *upper* limits of the class-intervals in which the scores have been tabulated. The PR of 74.5, for example, is 99.2; of 64.5, 92.0; of 44.5, 21.6, etc. These PR's are the ranks of given points in the distribution, and are not the PR's of scores.

(*b*) Percentiles and percentile ranks may also be determined quickly and fairly accurately from the ogive of the frequency distribution plotted in Figure 10. To obtain P_{50}, the median, for example, draw a line from 50 on the Y-scale parallel to the X-*axis* and where this line cuts the curve drop a perpendicular to the X-*axis*. This operation will locate the median at 51.5, approximately. The exact median, calculated from Table 14, is 51.65. Q_1 and Q_3 are found in the same way as the median. P_{25} or Q_1 falls approximately at 45.0 on the X-*axis*, and P_{75} or Q_3 falls at 57.0. These values may be compared with the calculated Q_1 and Q_3 which are 45.56 and 57.19, respectively. Other percentiles are read in the same way. To find P_{62}, for instance, begin with 62 on the Y-*axis*, go horizonally over to the curve, and drop a perpendicular to locate P_{62} approximately at 54.

In order to read the percentile rank of a given score from the ogive, we reverse the process followed in determining per-

centiles. Score 71, for example, has a *PR* of 97, approximately (see Figure 10). Calculation consists in starting with score 71 on the *X-axis*, going vertically up to the ogive, and horizontally across to the *Y-axis* to locate the *PR* at 97 on the cumulative percentage scale. The *PR* of score 47 is found in the same way to be approximately 30.

It will be noted that percentiles and percentile ranks are usually slightly in error when read from an ogive. If the curve is carefully drawn, however, the diagram fairly large and the scale divisions precisely marked, percentiles and *PR*'s may be read to a degree of accuracy sufficient for most purposes.

8. Other Uses of the Ogive

(1) Comparison of Groups

A useful over-all comparison of two or more groups is provided when ogives representing their scores on a given test are plotted upon the same coördinate axes. An illustration is given in Figure 11 which shows the ogives of the scores earned by two groups of children — 200 ten-year-old boys and 200 ten-year-old girls — upon an arithmetic reasoning test of sixty items. Data from which these ogives were constructed are given in Table 15.

Several interesting observations can be made from Figure 11. The boys' ogive lies to the right of the girls' over the entire range, showing that the boys score consistently higher than the girls. Differences in achievement as between the two groups are shown by the distances separating the two curves at various levels. It is clear that differences at the extremes — between the very high-scoring and the very low-scoring boys and girls — are not so great as are differences over the middle range. A more detailed analysis of the achievement of these two groups comes out in a comparison of certain points in the distribution. The boys' median is approximately 42, the girls' 32; and the difference between these measures is represented in Figure 11 by the line *AB*. The difference between the boys' Q_1 and the girls' Q_1 is represented by the line *CD*; and the difference be-

tween the two Q_3's is shown by the line EF. It is clear that the groups differ more at the median than at either quartile, and are farther separated at Q_3 than at Q_1.

TABLE 15

FREQUENCY DISTRIBUTIONS OF THE SCORES MADE BY 200 TEN-YEAR-OLD BOYS AND 200 TEN-YEAR-OLD GIRLS ON AN ARITHMETIC REASONING TEST

Scores	Boys f	Cum. f	Cum. % f	Smoothed Cum. Percentage f	Girls f	Cum. f	Cum. % f	Smoothed Cum. Percentage f
60–64	0	200	100.0	100.0	0	200	100.0	100.0
55–59	2	200	100.0	99.7	1	200	100.0	99.8
50–54	25	198	99.0	95.2	0	199	99.5	99.7
45–49	48	173	86.5	82.7	9	199	99.5	98.0
40–44	47	125	62.5	62.7	27	190	95.0	92.0
35–39	19	78	39.0	43.7	44	163	81.5	78.7
30–34	26	59	29.5	28.3	43	119	59.5	59.7
25–29	15	33	16.5	18.3	40	76	38.0	38.5
20–24	9	18	9.0	10.0	10	36	18.0	23.0
15–19	7	9	4.5	4.8	20	26	13.0	12.0
10–14	2	2	1.0	1.8	1	6	3.0	6.2
5–9	0	0	0	.3	2	5	2.5	2.3
0–4	0	0	0	0	3	3	1.5	1.3
	200				200		0	.5

$$\text{Rate} = \frac{1}{200} = .005$$

The extent to which one distribution overlaps another, whether at the median or at other designated points, can be determined quite readily from their ogives. By extending the vertical line through B (the boys' median) up to the ogive of the girls' scores, it is clear that approximately 88% of the girls fall below the boys' median. Hence, approximately 12% of girls exceed the median of the boys in arithmetic reasoning. Computing overlap from boys to girls, we find that approximately 76% of the boys exceed the girls' median. The vertical line through A (girls' median) cuts the boys' ogive at approximately the 24th percentile. Therefore 24% of the boys fall below the girls' median, and 76% are above this point. Still another illustration may be helpful. Suppose the problem is to determine what percentage of the girls score at or above the boys' 60th percentile. The answer is found by locating first the point

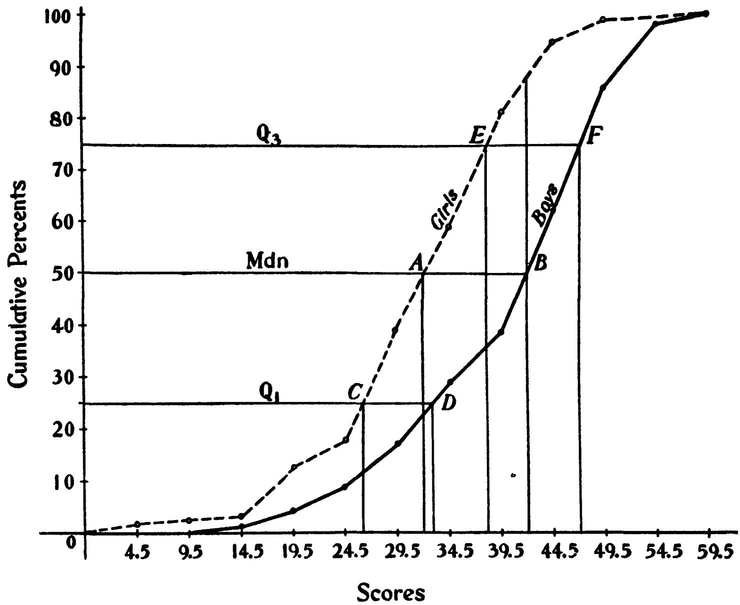


FIG. 11. Ogives Representing Scores Made by 200 Boys and 200 Girls on an Arithmetic Reasoning Test. (See Table 15.)

where the horizontal line through 60 cuts the boys' ogive. We then find the point on the girls' ogive directly above this value, and from here proceed horizontally across to locate the percentile rank of this point at 93. Since 93% of the girls fall below the boys' 60th percentile, about 7% score above this point.

(2) Percentile Norms

Norms are measures of achievement which represent the typical performance of a designated group or groups. The norm for ten-year-old boys in height, and the norm for seventh-grade pupils in City X in arithmetic is usually the mean or the median for the group. But norms may be much more detailed and may be reported for other points in the distribution as, for example, Q_1, Q_3, and various percentiles.

Percentile norms are especially useful in dealing with educational achievement examinations, when one wishes to evaluate and compare the achievement of a given student in a number of subject-matter tests. If the student earns a score of 63 on an achievement test in arithmetic, and a score of 143 on an achievement test in English, we have no way of knowing from the scores alone whether his achievement is good, medium, or poor, or how his standing in arithmetic and in English compare. If, however, we know that a score of 63 in arithmetic has a *PR* of 52, and a score of 143 in English a *PR* of 68, we may say at once that this student is average in arithmetic (52% of the students score lower than he) and good in English (68% score below him).

Percentile norms may be determined directly from the smoothed ogives of score distributions. Figure 12 represents the smoothed ogives of the two distributions of scores in arithmetic reasoning given in Table 15. Vertical lines drawn to the base line from points on the ogive locate the various percentile points. In Table 16 below, selected percentile norms in the arithmetic reasoning test have been tabulated for boys and girls separately. This table of norms may, of course, be ex-

TABLE 16

PERCENTILE NORMS FOR ARITHMETIC REASONING TEST (TABLE 15) OBTAINED FROM SMOOTHED OGIVES IN FIGURE 12.

	Girls		Boys	
Cum. %'s	Ogive	Calculated	Ogive	Calculated
99	52.0	49.0	57.5	54.5
95	46.5	44.5	54.5	52.9
90	43.5	42.7	52.5	50.9
80	40.0	39.2	49.0	48.1
70	37.0	36.9	46.5	46.1
60	35.0	34.6	44.0	44.0
50	32.5	32.5	41.5	41.8
40	30.0	30.0	39.0	39.7
30	27.0	27.5	35.0	34.8
20	23.5	25.0	30.0	30.9
10	18.5	18.0	24.5	25.2
5	14.0	15.5	19.5	20.1
1	3.5	3.3	6.5	14.5

tended by the addition of other intermediate or extreme values. Calculated percentiles are included in the table for comparison with percentiles read from the smoothed ogives. These calculated values are useful as a check on the graphically determined points, but ordinarily need not be found.

It is evident that percentile norms read from an ogive are not strictly accurate, but the error is slight except at the top

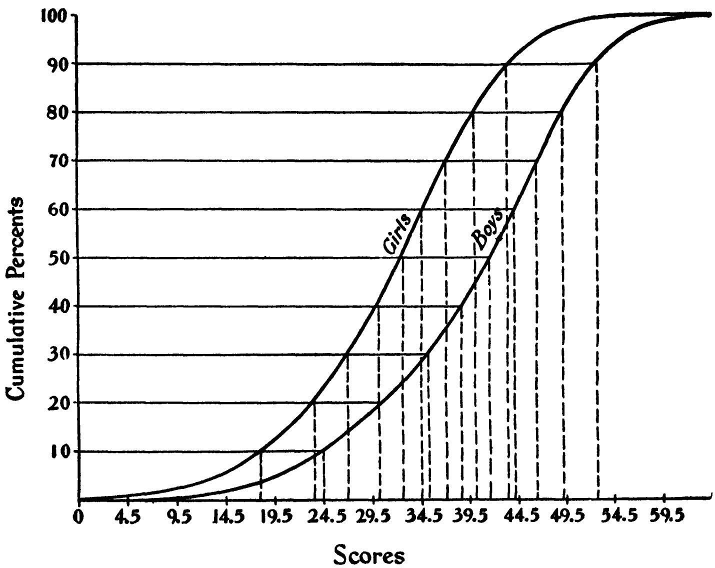


FIG. 12. Smoothed Ogives of the Scores in Table 15.

and bottom of the distribution. Estimates of these extreme percentiles from smoothed ogives are probably more nearly true values than are the calculated points, since the smoothed curve represents what we might expect to get from larger groups or in additional samplings.

The ogives in Figure 12 were smoothed in order to iron out minor kinks and irregularities in the curves. Owing to the smoothing process, these curves are more regular and continuous than are the original ogives in Figure 11. The only

difference between the process of smoothing an ogive and smoothing a frequency polygon (p. 16) is that we average cumulative percentage frequencies in the ogive instead of actual frequencies. Smoothed percentage frequencies are given in Table 15. The smoothed cumulative percent frequency to be plotted above 24.5, boys' distribution, is $\frac{16.5 + 9.0 + 4.5}{3}$, or 10.0; for the same point, girls' distribution, it is $\frac{38.0 + 18.0 + 13.0}{3}$ or 23.0. Care must be taken at the extremes of the distribution where the procedure is slightly different. In the boys' distribution, for example, the smoothed cumulative percent frequency at 9.5 is $\frac{1.0 + 0.0 + 0.0}{3}$ or .3%, and at 59.5, it is $\frac{100.0 + 100.0 + 99.0}{3}$ or 99.7. At 4.5 and 64.5, both of which lie outside the boys' distribution, the cumulative percentage frequencies are 100 $\left[\frac{100 + 100 + 100}{3}\right]$ and 0 $\left[\frac{0 + 0 + 0}{3}\right]$, respectively. Note that the smoothed ogive extends one interval beyond the original at both extremes of the distribution.

There is little justification for smoothing an ogive which is already quite regular or an ogive which is very jagged and irregular. In the first instance, smoothing accomplishes little if anything; in the second, it may seriously mislead. A smoothed curve shows what we might expect to get if the test or sampling, or both, were different (and perhaps better) than they actually were. Smoothing should never be a substitute for getting additional data or for constructing an improved test. It should certainly be avoided when the group is small and the ogive very irregular. Smoothing is perhaps most useful when the ogives show small irregularities here and there (see Figure 11) which may reasonably be assumed to have arisen from small and not very important factors.

IV. Other Graphical Methods

Data obtained from many problems in mental measurement, especially those which involve the study of changes attributable to growth, practice, learning, and fatigue, may be treated profitably by graphical methods. Two widely used devices are the *line graph*, frequently found in experimental psychology, and the *bar diagram* more often met with, perhaps, in education.

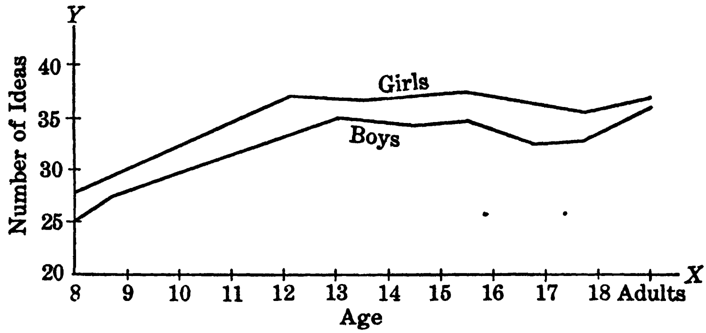


FIG. 13. Logical Memory. Age is represented on *X-line* (horizontal); Score, i.e., number of ideas remembered, on *Y-line* (vertical). (After Pyle.)

These two methods will be described in this section. For a discussion of other graphical methods, the reader is referred to books dealing specifically with the subject of graphics.*

1. The Line Graph

Figure 13 shows an age-progress curve. This graph represents the change in "logical memory for a connected passage" in boys and girls from eight to eighteen years old. Norms for adults are also included on the diagram. Age is represented on the horizontal or *X-axis* and "average number of ideas reproduced" at each age level is marked off on the vertical or

* For a simple treatment see Rugg, H. O., *A Primer of Graphics and Statistics for Teachers*, 1925. More advanced treatments may be found in Williams, J. H., *Graphic Methods in Education*, 1924, and Karsten, K. G., *Charts and Graphs*, 1923.

Y-axis. Memory ability as measured by this test rises to a peak at year 15 for both groups after which there is a slight decline followed by a rise at the adult level. There is a small but consistent sex difference throughout, the girls being higher on the average at each age.

Figure 14 illustrates the learning or practice curve. These curves show the improvement, in sending and receiving telegraphic messages, resulting from successive trials at the same

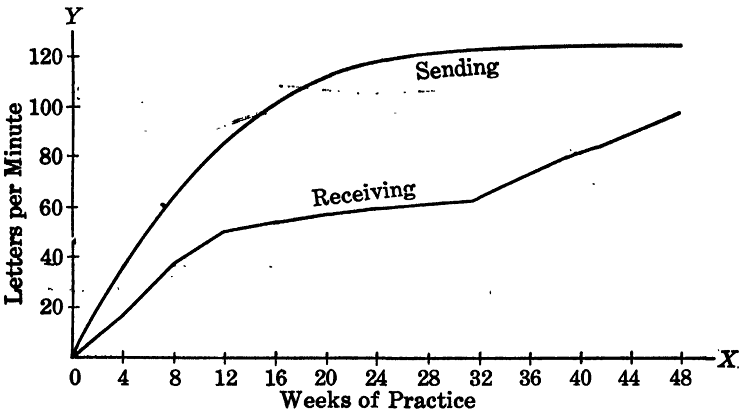


FIG. 14. Improvement in Telegraphy. Weeks of practice on *X-line;* number of letters per minute on *Y-line.* (After Bryan & Harter.)

task over a period of forty-eight weeks. Improvement as measured by the number of letters sent or received per minute is indicated along the *Y-axis*. Weeks of practice at the given task are represented by equal intervals on the *X-axis*.

Figure 15 is a performance or practice "curve." It represents twenty-five successive trials with the hand dynamometer made by one man and one woman. A marked sex difference in strength of grip is apparent throughout the practice period. Also as the experiment progressed a tendency to fatigue is evident in both subjects.

Figure 16 is Ebbinghaus' well-known "curve of retention."

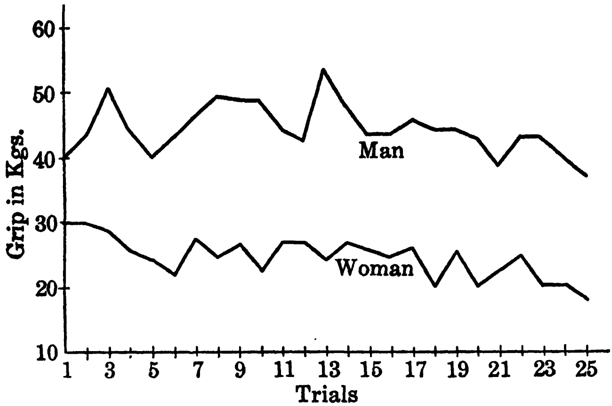


FIG. 15. Hand Dynamometer Readings in Kilograms for Twenty-five Successive Grips at Intervals of Ten Seconds. Two subjects, a man and a woman.

This curve represents memory retention as measured by the percentage of the original material retained after the passage of different time intervals. The time intervals between learning and relearning are laid off on the *X-axis;* and the percent retained, as measured by relearning, on the *Y-axis.*

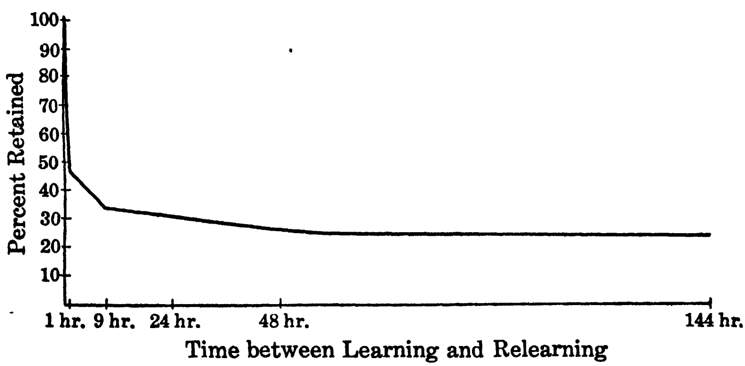


FIG. 16. Curve of Retention. The numbers on the baseline give hours elapsed from time of learning; numbers along *Y-axis* give percent retained.

2. The Bar Diagram

The bar graph is sometimes used in psychology to compare the relative amounts of some attribute (height, intelligence, educational achievement, etc.) possessed by two or more groups. In education the bar graph may be used to compare (usually in percentage terms) several different variables. Examples

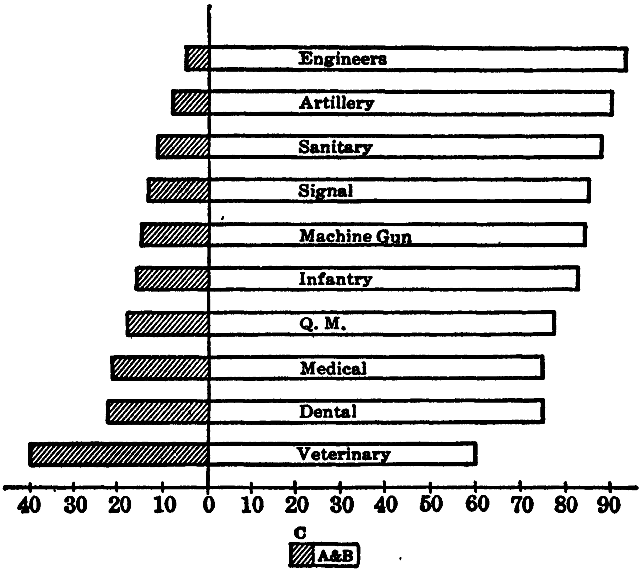


FIG. 17. Comparative Bar Graphs. The bars represent the percentage in each division of the military service receiving *A*'s and *B*'s or *C*'s.

are: the cost of instruction in various schools or in different counties; distribution of student time in and out of school; teachers' salaries by states or districts; relative expenditures for various purposes. The commonest form of the bar graph is that in which a set of bars is used, the lengths of the bars being proportional to the amounts of the variable possessed. For emphasis, a space is usually left between the bars, which are drawn side by side and may be either vertical or horizontal.

A horizontal bar graph is shown in Figure 17. These bars

represent the percentage of officers in various branches of the military service during World War I who received grades of *A* and *B* or *C* upon the Army Alpha Examination. The bars are arranged in order, the group receiving the highest percent of *A*'s and *B*'s being placed at the top. It is clear from the diagram that the Engineers, who ranked first, received about 95% *A*'s and *B*'s and about 5% *C*'s. The Veterinary Corps, which ranked last, received about 60% *A*'s and *B*'s and 40% *C*'s.

Another illustration of a bar graph is shown in Figure 18. The two parallel rectangles or "bars" represent student enrollment in two city high schools. Each bar is divided into four parts to represent freshmen, sophomores, juniors, and seniors. The size of a division is proportional to the percentage which each class is of the whole group. This type of graph is often called a *divided-bar graph.*

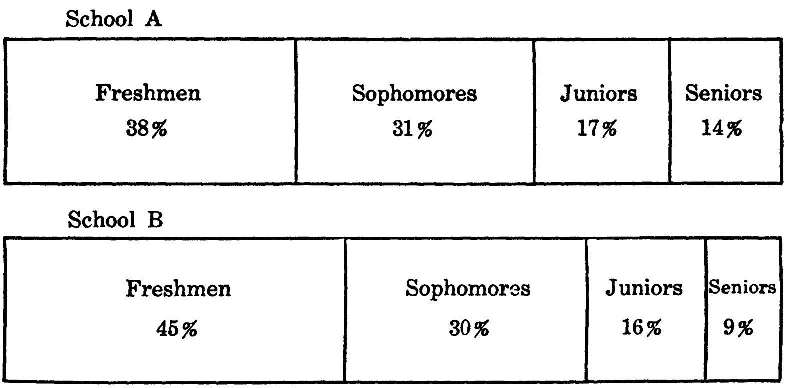


FIG. 18. Divided Bar Graphs. The two bars represent student enrollment in two high schools. Each bar is divided into four divisions. The length of a division show the proportion or percentage of students in that class.

PROBLEMS

1. The following distributions represent the achievement of two groups, A and B, upon a memory test.
   (*a*) Plot cumulative frequency graphs of Group *A*'s and of Group *B*'s scores, observing the 75% rule.
   (*b*) Plot ogives of the two distributions *A* and *B* upon the same axes.

(c) Determine P_{30}, P_{60}, and P_{90} graphically from each of the ogives and compare graphically determined with calculated values.
(d) What is the percentile rank of score 55 in Group *A*'s distribution? In Group *B*'s distribution?
(e) A percentile rank of 70 in Group *A* corresponds to what percentile rank in Group *B*?
(f) What percent of Group *A* exceeds the median of Group *B*?

Scores	Group A	Group B
79–83	6	8
74–78	7	8
69–73	8	9
64–68	10	16
59–63	12	20
54–58	15	18
49–53	23	19
44–48	16	11
39–43	10	13
34–38	12	8
29–33	6	7
24–28	3	2
	$N = 128$	$N = 139$

2. Construct an ogive of the following distribution of scores.

Scores	*f*
159.5 to 169.5	1
149.5 to 159.5	5
139.5 to 149.5	13
129.5 to 139.5	45
119.5 to 129.5	40
109.5 to 119.5	30
99.5 to 109.5	51
89.5 to 99.5	48
79.5 to 89.5	36
69.5 to 79.5	10
59.5 to 69.5	5
49.5 to 59.5	1
	$N = 285$

Read off percentile norms for the cumulative percentages:
99, 95, 90, 80 70, 60, 50, 40, 30, 20, 10, 5, and 1.

3. (*a*) In accordance with their scores upon a learning test, twenty children are ranked in order of merit. Calculate the percentile rank of each child.
   (*b*) If sixty children are ranked in order of merit, what is the percentile rank of the first, tenth, fortieth, and sixtieth?
4. Given the following data from five cities in the United States, represent the facts graphically by means of a bar graph.

	Percent of population which is		
City	Native White	Foreign-born White	Negro
A	.65	.30	.05
B	.60	.10	.30
C	.50	.45	.05
D	.40	.20	.40
E	.30	.10	.60

Answers

	Group *A*		Group *B*	
	Ogive	Cal.	Ogive	Cal.
1. (*c*) P_{30}	46.0	45.81	48.5	48.69
P_{60}	56.0	55.77	59.75	59.85
P_{90}	74.0	73.64	75.5	74.81

(*d*) 59; 49
(*e*) 62 (*f*) 39–40% of Group *A* exceed the median of Group *B*.

2. Read from ogive:

Cum. Percents:	99	95	90	80	70	60	50	40	30
Percentiles:	159	142.5	137.5	131.5	124.5	116.5	107	102	96.5
	20	10	5	1					
	91	82.5	79	64.5					

3. (*a*) 97.5; 92.5; 87.5; 82.5; 77.5; 72.5; 67.5; 62.5; 57.5; 52.5; 47.5; 42.5; 37.5; 32.5; 27.5; 22.5; 17.5; 12.5; 7.5; 2.5.
   (*b*) 99.17; 84.17; 34.17; .83.

Additional Problems and Questions on Chapters I–IV

1. Describe the characteristics of those distributions for which the mean is not an adequate measure of central tendency.
2. When is it inadvisable to use the coefficient of variation?
3. What is a multimodal distribution?

4. A student writes in a theme that by the application of eugenics it would be possible to raise the intelligence of the race, so that more people would be above the median I.Q. of 100. Comment on this statement.
5. Why cannot the σ of one test usually be compared directly with the σ of another test?
6. What effect will an increase in N probably have upon Q? (p. 54.)
7. What is the difference between a percentile and the ordinary percent grade used in school?
8. Does a percentile rank of 65 earned by a given pupil mean that 65% of the group make scores above him; that 65% make the same score; or that 65% make scores below him?
9. What is indicated by the relatively "flat" portion of an ogive?
10. Will increasing the size of the class-intervals used in grouping tend to make the frequency polygon more irregular?
11. Calculate the mean, median, mode, Q, and SD for each of the following distributions:

(1) Scores	f	(2) Scores	f	(3) Scores	f
90–99	2	14–15	3	25	1
80–89	12	12–13	8	24	2
70–79	22	10–11	15	23	6
60–69	20	8–9	20	22	8
50–59	14	6–7	10	21	5
40–49	4	4–5	4	20	2
30–39	1		$N = 60$	19	1
	$N = 75$				$N = 25$

12. (*a*) Plot the distribution in 11 (1) as a frequency polygon and histogram upon the same coördinate axes.
    (*b*) Plot the distribution in 11 (2) as an ogive. Locate graphically the median, Q_1, and Q_3. Determine the PR of score 9; of score 12.

Answers

11. (1) Mean = 68.10; Median = 68.75; Mode = 70.05; Q = 9.01; SD = 12.50

    (2) Mean = 9.23; Median = 9.10; Mode = 8.84; Q = 1.69; SD = 2.48

(3) Mean = 22.04
Median = 22.06
Mode = 22.10
Q = .91
SD = 1.34

12. (*b*) $Mdn = 9.0$; $Q_1 = 7.5$; $Q_3 = 11.0$ (Read from ogive)
PR of 9 = 50; of 12 = 84.5

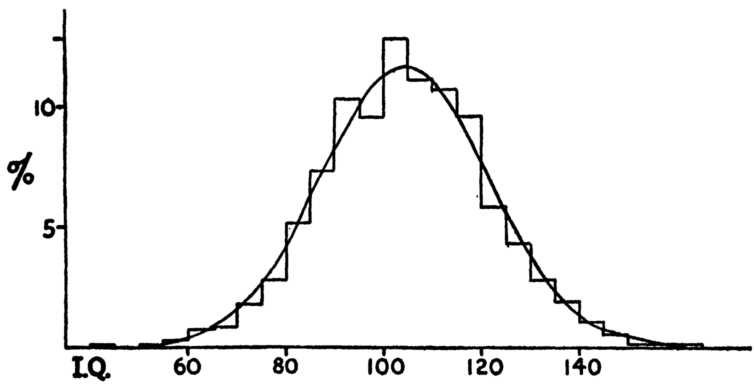
CHAPTER V

THE NORMAL PROBABILITY CURVE

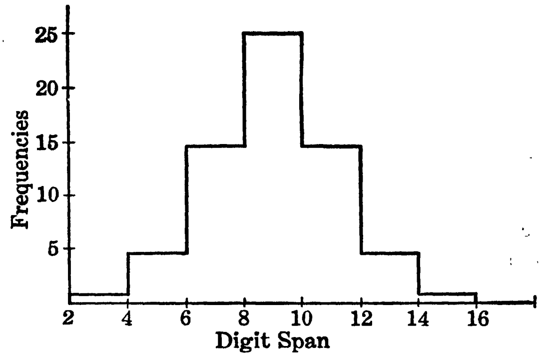
I. The Meaning and Importance of the Normal Probability Distribution

1. Introduction

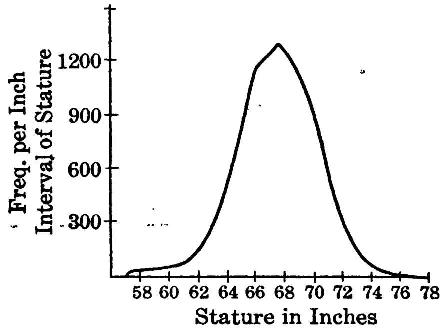
In Figure 19 are four diagrams, two polygons and two histograms, which represent frequency distributions of data drawn from anthropometry, psychology, and meteorology. It is apparent, even upon superficial examination, that all of these graphs have the same general form — the measures are concentrated closely around the center and taper off from this central high point or crest to the left and right. There are relatively few measures at the "low-score" end of the scale; an increasing number up to a maximum at the middle position; and a progressive falling-off toward the "high-score" end of the scale. If we divide the area *under* each curve (the area between the curve and the *X-axis*) by a line drawn per-



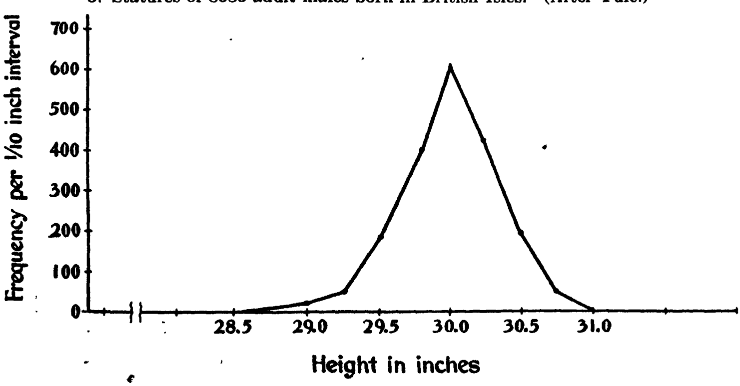
1. Form L I.Q. distribution and best-fitting normal curve, ages 2½ to 18.
(*from* McNemar, Quinn, *The Revision of the Stanford-Binet Scale*, p. 19)



2. Memory span for digits, 123 adult women students. (After Thorndike.)



3. Statures of 8585 adult males born in British Isles. (After Yule.)



4. Frequency distribution of barometer heights at Southampton: 4748 observations. (After Yule.)

FIG. 19. Frequency Distributions Drawn from Different Fields.

pendicularly through the central high point to the baseline, the two parts thus formed will be similar in shape and very nearly equal in area. It is clear, therefore, that each figure exhibits almost perfect bilateral symmetry. The perfectly symmetrical curve, or frequency surface, to which all of the graphs in Figure 19 approximate, is shown in Figure 20. This bell-shaped figure is called the *normal probability curve,* or simply the *normal curve,* and is of great value in mental measurement. An understanding of the characteristics of the

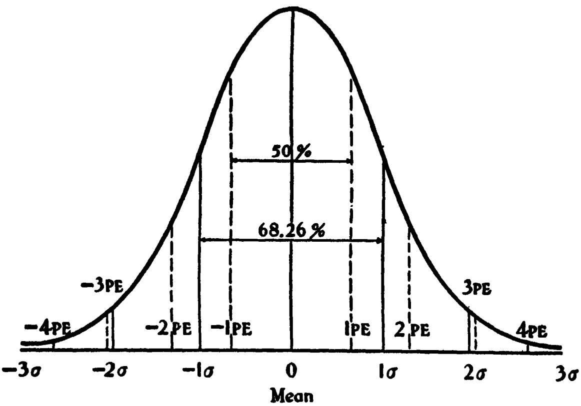


FIG. 20. Normal Probability Curve.

frequency distribution represented by the normal curve is essential to the student of experimental psychology and mental measurement. This chapter, therefore, will be concerned with the normal distribution, and its frequency polygon, the normal probability curve.

2. Elementary Principles of Probability

Perhaps the simplest approach to an understanding of the normal probability curve is through a consideration of the elementary principles of probability. As used in statistics, the "probability" of a given event is defined as the expected frequency of occurrence of this event among events of a like sort. This expected frequency of occurrence may be based

upon a knowledge of the conditions determining the occurrence of the phenomenon, as in dice-throwing or coin-tossing, or upon empirical data, as in mental and social measurements.

The probability of an event may be stated most simply, perhaps, as a ratio. We know, for example, that the probability of an unbiased coin falling heads is 1/2, and that the probability of a die showing a two-spot is 1/6. These ratios, called *probability ratios*, are defined by that fraction the numerator of which equals the desired outcome or outcomes and the denominator of which equals the total possible outcomes. A probability ratio always falls between the limits .00 (impossibility of occurrence) and 1.00 (certainty of occurrence). Thus the probability that the sky will fall is .00; that an individual now living will some day die is 1.00. Between these limits are all possible degrees of likelihood which may be expressed by appropriate ratios.

Let us now apply these simple principles of probability to the specific case of what happens when we toss coins.* If we toss one coin, obviously it must fall either heads (H) or tails (T) 100% of the time; and furthermore, since there are only two possible outcomes, a head or a tail is *equally probable*. Expressed as a ratio, therefore, the probability of H is 1/2; of T 1/2; and

$$(H + T) = 1/2 + 1/2 = 1.00$$

If we toss two coins, (*a*) and (*b*), at the same time, there are four possible arrangements which the coins may take:

(1)	(2)	(3)	(4)
a b	*a b*	*a b*	*a b*
H H	H T	T H	T T

Both coins (*a*) and (*b*) may fall H; (*a*) may fall H and (*b*) T; (*b*) may fall H and (*a*) T; or both coins may fall T. Expressed as ratios, the probability of *two* heads is 1/4 and the probability

* Coin-tossing and dice-throwing furnish easily understood and often used illustrations of the so-called "laws of chance."

of *two* tails 1/4. Also, the probability of an HT combination is 1/4, and of a TH combination 1/4. And since it ordinarily makes no difference which coin falls H or which falls T, we may add these two ratios (or double the one) to obtain 1/2 as the probability of an HT combination. The sum of our probability ratios is $1/4 + 1/2 + 1/4$ or 1.00.

Let us go a step farther and increase the number of coins to three. If we toss three coins (*a*), (*b*), and (*c*) simultaneously, there are eight possible outcomes:

(1)			(2)			(3)			(4)			(5)			(6)			(7)			(8)		
a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
H	H	H	H	H	T	H	T	H	T	H	H	H	T	T	T	H	T	T	T	H	T	T	T

Expressed as ratios, the probability of *three* heads is 1/8 (combination 1); of *two* heads and *one* tail 3/8 (combinations 2, 3, and 4); of *one* head and *two* tails 3/8 (combinations 5, 6, and 7); and of *three* tails 1/8 (combination 8). The sum of these probability ratios is $1/8 + 3/8 + 3/8 + 1/8$ or 1.00.

By exactly the same method used above for two and for three coins, we can determine the probability of different combinations of heads and tails when we have four, five, or any number of coins. These various outcomes may be obtained in a somewhat more direct way, however, than by writing down all of the different combinations which may occur. If there are n independent factors, the probability of the presence or absence of each being the same, the "compound" probabilities of the appearance of various combinations of factors will be expressed by expansion of the binomial $(p + q)^n$. In this expression p equals the probability that a given event will happen, q the probability that the event will not happen, and the exponent n indicates the number of factors (e.g., coins) operating to produce the final result.* If we substitute H for p and T for q (tails = non-heads), we have for two coins

* We may, for example, consider our coins to be independent factors, the occurrence of a head to be the *presence* of a factor and the occurrence of a tail the *absence* of a factor. Factors will then be "present" or "absent" in the various heads-tails combinations.

$(H + T)^2$; and squaring, the binomial $(H + T)^2 = H^2 + 2HT + T^2$. This expansion may be written,

1 H^2	1 chance in 4 of 2 heads; *probability ratio*	= 1/4
2 HT	2 chances in 4 of 1 head and 1 tail; *probability ratio*	= 1/2
1 T^2	1 chance in 4 of two tails; *probability ratio*	= 1/4
Total = 4		

These outcomes are identical with those obtained above by listing the three different combinations possible when two coins are tossed.

If we have three independent factors operating, the expression $(p + q)^n$ becomes for three coins $(H + T)^3$. Expanding this binomial, we get $H^3 + 3H^2T + 3HT^2 + T^3$, which may be written,

1 H^3	1 chance in 8 of 3 heads; *probability ratio*	= 1/8
3 H^2T	3 chances in 8 of 2 heads and 1 tail; *probability ratio*	= 3/8
3 HT^2	3 chances in 8 of 1 head and 2 tails; *probability ratio*	= 3/8
1 T^3	1 chance in 8 of 3 tails; *probability ratio*	= 1/8
Total = 8		

Again these results are identical with those got by listing the four different combinations possible when three coins are tossed.

The binomial expansion may be applied still more generally to those cases in which there are a larger number of independent factors operating. If we toss ten coins simultaneously, for instance, we have by analogy with the above, $(p + q)^{10}$. This expression may be written $(H + T)^{10}$, H standing for the probability of a head, T for the probability of a non-head (tail), and 10 for the number of coins tossed. When the binomial $(H + T)^{10}$ is expanded, the terms are

$$H^{10} + 10H^9T + 45H^8T^2 + 120H^7T^3 + 210H^6T^4 + 252H^5T^5 + 210H^4T^6 + 120H^3T^7 + 45H^2T^8 + 10HT^9 + T^{10}$$

which may be summarized as follows:

		Probability Ratio
1 H^{10}	1 chance in 1024 of all coins falling heads	$\frac{1}{1024}$
10 H^9T^1	10 chances in 1024 of 9 heads and 1 tail...	$\frac{10}{1024}$
45 H^8T^2	45 chances in 1024 of 8 heads and 2 tails..	$\frac{45}{1024}$
120 H^7T^3	120 chances in 1024 of 7 heads and 3 tails..	$\frac{120}{1024}$
210 H^6T^4	210 chances in 1024 of 6 heads and 4 tails..	$\frac{210}{1024}$
252 H^5T^5	252 chances in 1024 of 5 heads and 5 tails..	$\frac{252}{1024}$
210 H^4T^6	210 chances in 1024 of 4 heads and 6 tails..	$\frac{210}{1024}$
120 H^3T^7	120 chances in 1024 of 3 heads and 7 tails..	$\frac{120}{1024}$
45 H^2T^8	45 chances in 1024 of 2 heads and 8 tails..	$\frac{4}{1024}$
10 HT^9	10 chances in 1024 of 1 head and 9 tails...	$\frac{10}{1024}$
1 T^{10}	1 chance in 1024 of all coins falling tails..	$\frac{1}{1024}$
Total = 1024		

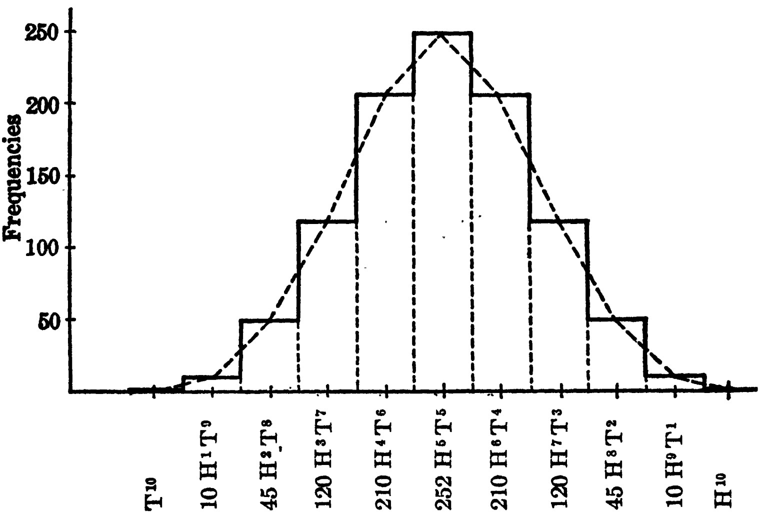


FIG. 21. Probability Surface Obtained from the Expansion of $(H + T)^{10}$.

These data are represented graphically in Figure 21 by a histogram and frequency polygon plotted on the same axes. The eleven terms of the expansion have been laid off at equal distances along the *X-axis*, and the "chances" of the occurrence of each combination of H's and T's are plotted as frequencies

on the *Y-axis*. The result is a symmetrical frequency polygon with the greatest concentration in the center and the "scores" falling away by corresponding decrements above and below the central high point. Figure 21 represents the results to be expected *theoretically* when ten coins are tossed 1024 times.

Many experiments have been conducted, in which coins were tossed or dice thrown a great many times, with the idea of checking theoretical against actual results. In one well-known experiment,* twelve dice were thrown 4096 times. Each four-, five-, and six-spot combination was taken as a "success" and

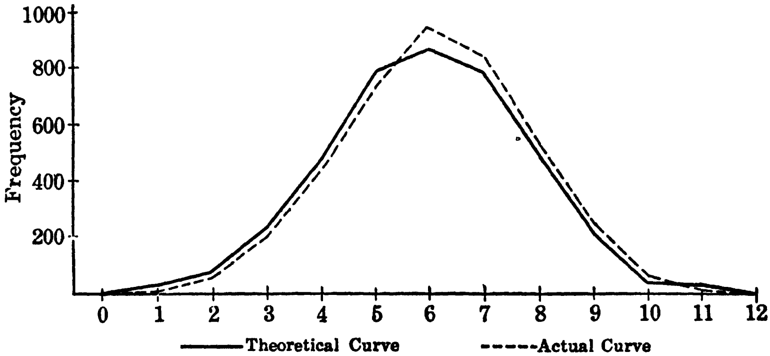


FIG. 22. Comparison of Observed and Theoretical Results in Throwing Twelve Dice 4096 Times. (After Yule.)

each one-, two-, and three-spot combination as a "failure." Hence the probability of success and the probability of failure were the same. In a throw showing the faces 3, 1, 2, 6, 4, 6, 3, 4, 1, 5, 2, and 3, there would be five successes and seven failures. The *observed* frequency of the different numbers of successes and the *theoretical* outcomes obtained from the expansion of the binomial expression $(p + q)^{12}$ have been plotted on the same axes in Figure 22. The reader will note that the observed frequencies correspond quite closely to the theoretical except for a tendency to shift slightly to the right. If, as an experiment, the reader will toss ten coins 1024 times his results

* Weldon's experiment; see Yule, G. U., *An Introduction to the Theory of Statistics* (10th ed., 1932), p. 258.

will be in close agreement with the theoretical outcomes shown in Figure 21.

Throughout the discussion in this section, we have taken the probability of occurrence (e.g., H) and the probability of non-occurrence (non-H or T) of a given factor to be the same. This is not a necessary condition, however. For instance, the probability of an event's happening may be only 1/5; of its not happening, 4/5. Any probability ratio is possible as long as $(p + q) = 1.00$. But distributions obtained from the expansion of $(p + q)^n$ when p is not equal to q are "skewed" or asymmetrical and are not normal (p. 129).

3. Use of the Probability Curve in Mental Measurement

The frequency curve plotted in Figure 21 from the expansion of the expression $(\mathrm{H} + \mathrm{T})^{10}$ is a symmetrical many-sided polygon. If the number of factors (e.g., coins) determining this polygon were increased from 10 to 20, to 30, and then to 100, say (the baseline extent remaining the same), the faces of the polygon would increase regularly in number from 23 to 203. With each increase in the number of factors, the faces of the figure would become shorter, and the points on the frequency surface would move closer together. Finally, when the number of factors became very large — when n in the expression $(p + q)^n$ became infinite — the polygon would exhibit a perfectly smooth surface like that of the curve in Figure 20. This "ideal" polygon or "normal" curve represents the frequency of occurrence of various combinations of a very large number of *equal, similar*, and *independent* factors (e.g., coins), when the probability of the appearance (e.g., H) or non-appearance (e.g., T) of each factor is the same.

If we compare the four graphs plotted from measures of height, intelligence, memory span, and barometric readings in Figure 19, with the normal probability curve in Figure 20, the similarity of these diagrams to the normal curve is clearly evident. The resemblance of these and many other distributions to the normal seems to express a general tendency of

quantitative data to take the symmetrical, bell-shaped form. This general tendency may be stated in the form of a "principle" as follows: measurements of many natural phenomena and of many mental and social traits under certain conditions *tend* to be distributed symmetrically about their means in proportions which approximate those of the normal probability distribution.

Much evidence has accumulated to show that the normal distribution serves to describe the frequency of occurrence of many variable facts with a relatively high degree of accuracy. Various phenomena which follow the normal probability curve (at least approximately) may be classified as follows:

1. *Biological statistics:* the proportion of male to female births for the same country or community over a period of years; the proportion of different types of plants and animals in cross-fertilization (the Mendelian ratios).

2. *Anthropometrical data:* height, weight, cephalic index, etc., for large groups of the same age and sex.

3. *Social and economic data:* rates of birth, marriage, or death under certain constant conditions; wages and output of large numbers of workers in the same occupation under comparable conditions.

4. *Psychological measurements:* intelligence as measured by standard tests; speed of association, perception-span, reaction-time; educational test scores, e.g., in spelling, arithmetic, reading.

5. *Errors of observation:* measures of height, speed of movement, linear magnitudes, physical and mental traits, and the like, contain errors which are as likely to cause them to deviate above as below their true values. Chance errors of this sort vary in magnitude and sign and occur in frequencies which follow closely the normal probability curve.*

It is an interesting speculation that many frequency distributions of scores and other measures are similar to those ob-

* This topic is treated in Chapter VII.

tained by tossing coins or throwing dice because the former, like the latter, are actually probability distributions. The symmetrical normal distribution, as we have seen, represents the probability of occurrence of the various possible combinations of a great many factors (e.g., coins). In a normal distribution all of the n factors are taken to be *similar*, *independent*, and *equal in strength;* and the probability that each will be present (e.g., show an H) or absent (e.g., show a T) is the same. The appearance on a coin of a head or a tail is undoubtedly determined by a large number of small (or "chance") influences as liable to work one way as another. The twist with which the coin is spun may be important, as well as the height from which it is thrown, the weight of the coin, the kind of surface upon which it falls, and many other circumstances of a like sort. By analogy, the presence or absence of each one of the large number of genetic factors which determine the shape of a man's head, or his intelligence, or his personality, may depend upon a host of adventitious influences whose net effect we call "chance."

But the striking similarity of obtained and probability distributions should not lead us to conclude that *all* distributions of mental and physical traits which exhibit a symmetrical form have *necessarily* arisen through the operation of those principles which govern the appearance of dice or coin combinations. The factors which determine musical ability, let us say, or mechanical skill are too little known to justify the assumption, *a priori*, that they combine in the same proportions as do the head and tail combinations in "chance" distributions of coins. Moreover, the psychologist usually constructs his tests with the normal hypothesis definitely in mind. The resulting symmetrical distribution is to be taken, then, as evidence of the success of his efforts rather than as conclusive proof of the "normality" of the trait being measured.*

The selection of the normal rather than some other type curve

* McNemar, Q., *The Revision of the Stanford-Binet Scale* (1942), Chapter II.

is sufficiently warranted by the fact that this distribution generally does fit the data better, and is more useful. But the "theoretical justification and the empirical use of the normal curve are two quite different matters." *

II. Properties of the Normal Probability Distribution

1. The Equation of the Normal Curve

The equation of the normal probability curve reads

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \tag{17}$$

(equation of the normal probability curve)

in which

x = scores (expressed as deviations from the mean) laid off along the baseline or *X-axis.*

y = the height of the curve above the *X-axis*, i.e., the frequency of a given x-value or the number achieving a certain score.

The other terms in the equation are constants: —

N = number of cases.

σ = standard deviation of the distribution.

π = 3.1416 (the ratio of the circumference of a circle to its diameter).

e = 2.7183 (base of the Napierian system of logarithms).

When N and σ are known, it is possible from equation (17) to compute (1) the frequency (or y) of a given value x, i.e., the *number* of individuals making a certain *score;* and (2) the number, or percentage, of individuals scoring between two points, or above or below a given point in the distribution. But these calculations are rarely necessary, as tables are available from which this information may be readily obtained. A knowledge of these tables (Tables 17 and 18) is extremely valuable in the solution of a number of problems. For this reason it is very

* Jones, D. C., *A First Course in Statistics* (1921), p. 233.

desirable that the construction and use of Tables 17 and 18 be clearly understood.

2. Tables of Areas under the Normal Curve

(1) Areas in terms of σ as unit.

Table 17 gives the fractional parts of the total area under the normal curve found between the mean and ordinates (y's) erected at various distances from the mean. In Table 17 distances along the *X-axis* are measured in σ units (see Fig. 20). The total area under the curve (the number of scores in the distribution) is taken arbitrarily to be 10,000, because of the greater ease with which fractional parts of the total area may then be calculated.

The first column of the table, x/σ, gives distances in tenths of σ measured off on the baseline of the normal curve from the mean as origin. We have already learned that $x = X - M$, i.e., that x measures the deviation of a score X from M. If x is divided by σ, deviation from the mean is expressed in σ-units. Such σ-deviation scores are often called *standard scores*, or z-scores ($z = x/\sigma$). Distances from the mean in hundredths of σ are given by the headings of the columns. To find the number of cases in a normal distribution between the mean and the ordinate erected at a distance of 1σ from the mean, go down the x column until 1.0 is reached, and in the next column under .00 take the entry opposite 1.0, viz., 3413. This figure means that 3413 cases in 10,000, or 34.13% of the entire area of the curve lie between the mean and 1σ. Put more exactly, 34.13% of the cases in a normal distribution fall within the area bounded by the baseline of the curve, the ordinate erected at the mean, the ordinate erected at a distance of 1σ from the mean, and the curve itself (see Fig. 20). To find the percentage of the distribution between the mean and 1.57σ, say, go down the x/σ column to 1.5, then across horizontally to the column headed .07, and take the entry 4418. This means that in a normal distribution, 44.18% of the area (N) lie between the mean and 1.57σ.

TABLE 17

FRACTIONAL PARTS OF THE TOTAL AREA (TAKEN AS 10,000) UNDER THE NORMAL PROBABILITY CURVE, CORRESPONDING TO DISTANCES ON THE BASELINE BETWEEN THE MEAN AND SUCCESSIVE POINTS LAID OFF FROM THE MEAN IN UNITS OF STANDARD DEVIATION

Example: between the mean and a point $1.38\sigma \left(\frac{x}{\sigma} = 1.38\right)$ are found 41.62% of the entire area under the curve.

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0.7	2580	2611	2642	2673	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3290	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4383	4394	4406	4418	4429	4441
1.6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2.0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2.9	4981	4982	4982	4983	4984	4984	4985	4985	4986	4986
3.0	4986.5	4986.9	4987.4	4987.8	4988.2	4988.6	4988.9	4989.3	4989.7	4990.0
3.1	4990.3	4990.6	4991.0	4991.3	4991.6	4991.8	4992.1	4992.4	4992.6	4992.9
3.2	4993.129									
3.3	4995.166									
3.4	4996.631									
3.5	4997.674									
3.6	4998.409									
3.7	4998.922									
3.8	4999.277									
3.9	4999.519									
4.0	4999.683									
4.5	4999.966									
5.0	4999.997133									

TABLE 18

FRACTIONAL PARTS OF THE TOTAL AREA (TAKEN AS 10,000) UNDER THE NORMAL PROBABILITY CURVE, CORRESPONDING TO DISTANCES ON THE BASELINE BETWEEN THE MEAN AND SUCCESSIVE POINTS LAID OFF FROM THE MEAN IN UNITS OF *PE*

Example: between the mean and a point 1.55 *PE* $\left(\frac{x}{PE} = 1.55\right)$ from the mean are found 35.21% of the entire area under the curve.

$\frac{x}{PE}$	.00	.05	$\frac{x}{PE}$	.00	.05
0	0000	0135	3.0	4785	4802
.1	0269	0403	3.1	4817	4832
.2	0537	0670	3.2	4846	4858
.3	0802	0933	3.3	4870	4881
.4	1063	1193	3.4	4891	4900
.5	1320	1447	3.5	4909	4917
.6	1571	1695	3.6	4924	4931
.7	1816	1935	3.7	4937	4943
.8	2053	2168	3.8	4948	4953
.9	2281	2392	3.9	4957	4961
1.0	2500	2606	4.0	4965	4968
1.1	2709	2810	4.1	4972	4974
1.2	2909	3004	4.2	4977	4979
1.3	3097	3187	4.3	4981	4983
1.4	3275	3360	4.4	4985	4987
1.5	3442	3521	4.5	4988	4989
1.6	3597	3671	4.6	4990	4991
1.7	3742	3811	4.7	4992	4993
1.8	3876	3939	4.8	4994	4995
1.9	4000	4058	4.9	4995	4996
2.0	4113	4166	5.0	4996	4997
2.1	4217	4265	5.1	4997.1	4997.4
2.2	4311	4354	5.2	4997.7	4998
2.3	4396	4435	5.3	4998.2	4998.5
2.4	4473	4508	5.4	4998.6	4998.8
2.5	4541	4573	5.5	4999	4999.1
2.6	4603	4631	5.6	4999.2	4999.3
2.7	4657	4682	5.7	4999.4	4999.5
2.8	4705	4727	5.8	4999.54	4999.6
2.9	4748	4767	5.9	4999.65	4999.7

We have so far considered only σ-distances measured in the *positive* direction from the mean; that is, we have taken account only of the *right* half — the high-score end — of the normal curve. Since the curve is bilaterally symmetrical, the entries in Table 17 apply to σ-distances measured in the *negative* direc-

tion (to the *left*) as well as to those measured in the positive direction. To find the percentage of the distribution between the mean and -1.26σ, for instance, take the entry in the column headed .06, opposite 1.2 in the x/σ column. This entry (3962) tells us that 39.62% of the cases in the normal distribution fall between the mean and -1.26σ. The percentage of cases between the mean and -1σ is 34.13; and the reader will now be able to verify the statement made on page 59 that between the mean and $\pm 1\sigma$ are 68.26% of the cases in a normal distribution (see also Fig. 20).

While the normal curve does not actually meet the baseline until we are at infinite distances to the right and left of the mean, for practical purposes the curve may be taken to end at points -3σ and $+3\sigma$ distant from the mean. Table 17 shows that 4986.5 cases in the total 10,000 fall between the mean and $+3\sigma$; and 4986.5 cases will, of course, fall between the mean and -3σ. Therefore, 9973 cases in 10,000, or 99.73% of the entire distribution, lie within the limits -3σ and $+3\sigma$. By cutting off the curve at these two points, therefore, we disregard only .27 of 1% of the distribution, a negligible amount except in very large samples.

(2) Areas in Terms of *PE* as Unit.

Instead of σ the *PE* may be used as the unit of measurement in determining the area within given parts of the normal curve. Table 18 gives fractional parts of the total area under the normal curve found between the mean and ordinates erected at various *PE* distances from the mean. This table is read in exactly the same way as Table 17. To find, for instance, the number of cases between the mean and $1PE$ (or more accurately the ordinate erected at this point), we go down the x/PE column to 1.0 and opposite this entry in the next column headed .00 read 2500. (Fig. 20) Twenty-five percent of the cases in the distribution, therefore, lie between the mean and $1PE$. In like manner, 25% of the cases lie between the mean and $-1PE$; and it is clear that the middle 50% of a normal distribution

fall between $-1PE$ and $+1PE$ measured off from the mean (p. 54). Table 18 cannot be read in as fine units as Table 17, only tenths and .05ths PE divisions being given. If smaller divisions are desired linear interpolation can readily be made with little error.

Just as we usually disregard that part of a normal curve beyond the limits $\pm 3\sigma$, we ordinarily ignore that part of the curve beyond the limits $\pm 4PE$. There are 9930 (4965×2) cases in the total 10,000 between the mean and $\pm 4PE$ (Table 18). Hence, in cutting off the curve at $\pm 4PE$, we lose only .70 of 1% of the cases in the distribution.

There is little to choose as between Tables 17 and 18. Table 17 admits of easier interpolation but Table 18 is accurate enough, without interpolation, for most purposes. Table 17 is more often used in mental measurement.

3. Relationships among Constants of the Normal Probability Curve

In the normal probability curve, the mean, the median, and the mode all fall exactly at the midpoint of the distribution and are numerically equal. Since the normal curve is bilaterally symmetrical, all of the measures of central tendency must coincide at the middle of the distribution.

The measures of variability include certain constant fractions of the total area of the normal curve, which may be read from Tables 17 and 18. Between the mean and $\pm 1\sigma$ lie the middle two-thirds (approximately) of the cases in the normal distribution. Between the mean and $\pm 2\sigma$ are found 95% (approximately) of the distribution; and between the mean and $\pm 3\sigma$ are found 99.7% (approximately 100%) of the distribution. There are 68 chances (approximately) in 100 that a score will lie within $\pm 1\sigma$ from the mean in the normal distribution; there are 95 chances in 100 that it will lie within $\pm 2\sigma$ from the mean; and 99.7 chances in 100 that it will lie within $\pm 3\sigma$ from the mean.

As we have seen, $\pm 1PE$ mark off the middle 50% of the

cases, i.e., the 25% of the measures directly above, and the 25% directly below, the measure of central tendency. Furthermore, $\pm$ $2PE$ include 82.26% of the measures in the distribution; $\pm$ $3PE$, 95.70% of the measures in the distribution; and $\pm$ $4PE$, 99.30% of the measures in the distribution.

The following constant relations exist among the measures of variability:

$$1.\quad PE = .6745\sigma$$
$$2.\quad \sigma = 1.4826PE$$

These equations may be verified from the percents of area included by each. Thus, we find by interpolation in Table 17, that $.6745\sigma$ ($1PE$) includes the 25% of the distribution just above (or below) the mean; also, from Table 18, that $1.48PE$ includes the 34% of the distribution just above (or below) the mean. From these formulas it is evident why it was stated earlier (p. 59) that σ is always greater than $Q(PE)$.

III. Measuring Divergence from Normality

1. Skewness

In a frequency polygon or histogram, usually the first thing which strikes the eye is the symmetry or the lack of symmetry in the figure. In the normal curve the mean, the median, and the mode all coincide and there is perfect balance between the right and left halves of the figure. A distribution is said to be "skewed" when the mean, the median, and the mode fall at different points in the distribution, and the balance (or center of gravity) is shifted to one side or the other, to right or left. It is important to know (1) whether the skewness which often occurs in distributions of test scores and other measures represents a real divergence from the normal form; or (2) whether such divergence is the result of chance fluctuations, arising from temporary causes, and is not significant of real discrepance. The degree of displacement or skewness in a frequency distribution may be determined by the formula

$$Sk = \frac{3(\text{mean} - \text{median})}{\sigma} \tag{18}$$

(*a measure of skewness in a frequency distribution*)

In a normal distribution the mean equals the median and the skewness is 0. The more nearly the distribution approaches the

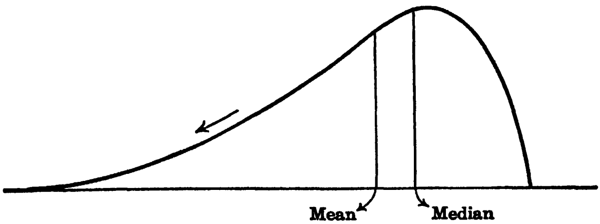


FIG. 23. Negative Skewness: to the Left.

normal form, the closer together are the mean and the median, and the less the skewness. Distributions are said to be skewed *negatively*, or to the *left*, when the scores are massed at the high end of the scale (the right end), and spread out gradually at the low or left end, as shown in Figure 23. Distributions are skewed *positively*, or to the *right*, when the scores are massed

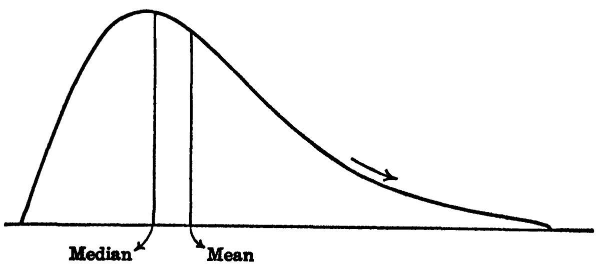


FIG. 24. Positive Skewness: to the Right.

at the low (the left) end of the scale, and spread out gradually toward the high or right end as shown in Figure 24.

If we apply formula (18) to the distribution of fifty Army Alpha scores in Table 1, page 6, −.28 is obtained as a measure of skewness. This result points to a slight negative skewness in the data, which may be seen by reference to Figure 2, page 14.

Formula (18) gives the measure of skewness for the distribution of the 200 cancellation scores (Table 3, p. 14) as .009. This negligible degree of positive skewness shows how closely this distribution approaches the symmetrical probability form.

Another measure of skewness is given by the formula

$$Sk = \frac{(P_{90} + P_{10})}{2} - P_{50} \tag{19}$$

(*a measure of skewness in terms of percentiles*) *

For the normal distribution Sk by formula (19) is zero: P_{50} lies just midway between P_{90} and P_{10}.

Applying this formula to the distributions of fifty Army Alpha scores and 200 cancellation scores, we obtain for the first $Sk = -2.50$; and for the second $Sk = .03$. These results are numerically different from the measures of skewness obtained from formula (18), because the two measures of skewness are computed from different reference values in the distribution, and hence are not directly comparable. The two formulas agree, however, in indicating some negative skewness for the distribution of fifty Alpha scores, and an insignificant degree of positive skewness for the 200 cancellation scores. In comparing the skewness of two distributions we should use either formula (18) or (19); not first the one and then the other.

The important question of how much skewness a distribution must exhibit before it may be said to be *significantly* skewed cannot be answered until we have calculated a "standard error" of our measure of skewness. A formula for the standard error of Sk, when determined by formula (19), and a method of testing whether the skewness of a given distribution is significant is discussed in Chapter VII, page 220.

2. Kurtosis

The term kurtosis refers to the "peakedness" or flatness of a frequency distribution as compared with the normal. A fre-

* Kelley, T. L., *Statistical Method* (1923), p. 77. The terms in this formula, as given by Kelley, have been reversed so that the sign of Sk will agree with the conventional notion of positive and negative skewness.

quency distribution more peaked than the normal is said to be *leptokurtic;* one flatter than the normal, *platykurtic.* Figure 25 shows a leptokurtic distribution and a platykurtic distribution plotted on the same diagram around the same mean. A normal curve (called *mesokurtic*) has also been drawn in on the

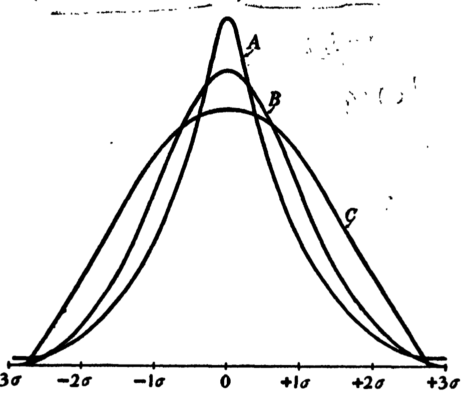


FIG. 25. Leptokurtic (A), Normal or Mesokurtic (B) and Platykurtic (C) Curves.

diagram to bring out the contrast in the figures, and to make comparison easier. A formula for measuring kurtosis is

$$Ku = \frac{Q}{(P_{90} - P_{10})} \tag{20}$$

(*a measure of kurtosis in terms of percentiles*)

For the normal curve, formula (20) gives $Ku = .263$.* If Ku is *greater* than .263 the distribution is platykurtic; if *less than* .263 the distribution is leptokurtic. Calculating the kurtosis of the distributions of fifty Alpha scores and 200 cancellation scores, discussed above, we obtain $Ku = .237$ for the first distribution, and $Ku = .223$ for the second. Both distributions, therefore, are slightly leptokurtic. To determine

* From Table 18, we find that $Q(PE) = 1.00$, $P_{90} = 1.90$ and $P_{10} = -1.90$. Hence, by formula (20)

$$Ku = \frac{1.00}{[1.90 - (-1.90)]} = \frac{1.00}{3.80} = .263$$

whether the kurtosis in a distribution is significant, that is, whether the curve is too high or too flat to be treated as sensibly normal, we must evaluate Ku in terms of its standard error. A formula for the standard error of Ku, and a method of determining the significance of an obtained measure of Ku will be given in Chapter VII, page 220.

3. Comparing a Given Histogram or Frequency Polygon with a Normal Curve of the Same Area, M and σ

In this section methods will be described for superimposing on a given histogram or frequency polygon a normal curve of the same N, M, and σ as the actual distribution. Such a normal curve is the "best fitting" normal distribution for the given data. The research worker often wishes to compare his distribution "by eye" with that normal curve which "best fits" the data, and such a comparison may profitably be made even if no measures of divergence from normality are computed. In fact, the direction and extent of asymmetry often strike us more convincingly when seen in a graph than when expressed by measures of skewness and kurtosis. It may be noted that a normal curve can always be readily constructed by following the procedures given here provided the area (N) and variability (σ) are known.

Table 19 shows the frequency distribution of scores made on the Thorndike Intelligence Examination by 206 college freshmen. The mean is 81.59, the median 81.00, and the σ 12.14. This frequency distribution has been plotted in Figure 26, and over it on the same axes has been drawn in the best fitting normal curve, i.e., the normal curve which best describes these data. The Thorndike scores are represented by a histogram instead of by a frequency polygon in order to prevent coincidence of the surface outlines and to bring out more clearly agreement and disagreement at different points. To plot a normal curve over this histogram, we first compute the height of the maximum ordinate (y_o) or the frequency at the middle of the distribution. The maximum ordinate (y_o) can be determined

TABLE 19

FREQUENCY DISTRIBUTION OF THE SCORES MADE BY 206 FRESHMEN ON THE THORNDIKE INTELLIGENCE EXAMINATION

Scores	f	
115–119	1	
110–114	2	
105–109	4	
100–104	10	Mean = 81.59
95–99	13	Median = 81.00
90–94	18	σ = 12.14
85–89	34	
80–84	30	
75–79	37	
70–74	27	
65–69	15	
60–64	10	
55–59	2	
50–54	2	
45–49	1	
	$N = 206$	

from the equation of the normal curve given on page 113. When x in this equation is put equal to zero (the x at the mean of the normal curve is 0), the term $e^{\frac{-x^2}{2\sigma^2}}$ equals 1.00, and $y_o = \frac{N}{\sigma\sqrt{2\pi}}\cdot$ In the present problem, $N = 206$; $\sigma = 2.43$ * (in units of class-interval), and $\sqrt{2\pi} = 2.51$; hence $y_o = 33.8$ (see Fig. 26 for calculations). Knowing y_o, we are able to compute from Table 20 the heights of ordinates at given distances from the mean. The entries in Table 20 give the heights of the ordinates in the normal probability curve, at various σ-distances from the mean, expressed as fractions of the maximum or middle ordinate taken equal to 1.00000. To find, for example, the height of the ordinate at $\pm 1\sigma$, we take the entry .60653 from the table opposite $x/\sigma = 1.0$. This means that when the maximum central ordinate (y_o) is 1.00000, the ordinate (i.e., frequency) $\pm 1\sigma$ removed from M is .60653; or the frequency at $\pm 1\sigma$ is about 61% of the maximum frequency at the middle of the distribution. In Figure 26 the ordinates $\pm 1\sigma$ from M

* $\sigma = 2.43 \times 5$ (interval). The σ in interval units is used in the equation, since the units on the *X-axis* are in terms of class-intervals.

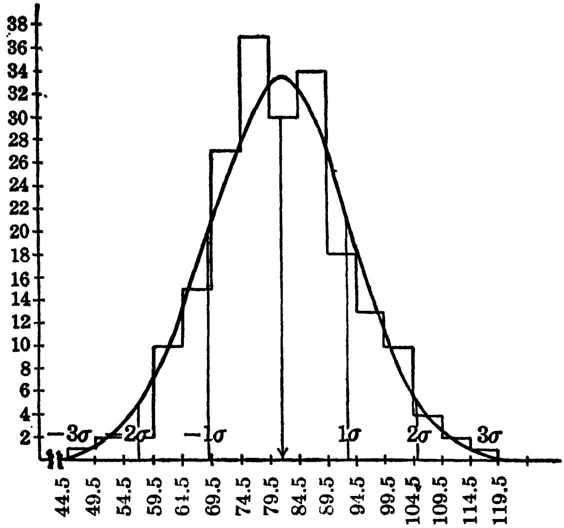


FIG. 26. Frequency Distribution of the Scores of 206 Freshmen on the Thorndike Intelligence Examination, Compared with Best-Fitting Normal Curve for Same Data. (For data, see Table 19.)

NORMAL CURVE ORDINATES AT MEAN, ± 1σ, ± σ2, ± 3σ

$$y_o = \frac{N}{\sigma\sqrt{2\pi}} = \frac{206}{2.43 \times 2.51} = 33.8$$

$$\begin{aligned} \pm 1\sigma &= .60653 \times 33.8 = 20.5 \\ \pm 2\sigma &= .13534 \times 33.8 = 4.6 \\ \pm 3\sigma &= .01111 \times 33.8 = .4 \end{aligned}$$

are $.60653 \times 33.8$ (y_o) or 20.5. The ordinates $\pm 2\sigma$ from M are $.13534 \times 33.8$ or 4.6; and the ordinates $\pm 3\sigma$ from M are $.01111 \times 33.8$ or .4.

The normal curve may be sketched in without much difficulty through the ordinates at these seven points. Somewhat greater accuracy may be obtained if various intermediate ordinates, for example, at $\pm .5\sigma$, $\pm 1.5\sigma$, etc., are also plotted. The ordinates for the curve in Figure 26 at $\pm .5\sigma$ are $.88250 \times 33.8$ or 29.3; at $\pm 1.5\sigma$, $.32465 \times 33.8$ or 11.0, etc.

From formula (19) the skewness of our distribution of 206 scores is found to be 1.25. This small value indicates a low degree of positive skewness in the data. The kurtosis of the

TABLE 20

ORDINATES OF THE NORMAL PROBABILITY CURVE EXPRESSED AS FRACTIONAL PARTS OF THE MEAN ORDINATE, y_o

The height of the ordinate erected at the mean can be computed from $y_o = \frac{N}{\sigma\sqrt{2\pi}}$ where $\sqrt{2\pi} = 2.51$. The height of any other ordinate, in terms of y_o, can be read from the table when one knows the distance which the ordinate is from the mean. For example: the height of an ordinate at a distance of 1.50σ from the mean is $.32465\ y_o$; the height of an ordinate at a distance of $-\ 2.37\sigma$ from the mean is $.06029\ y_o$.

$\frac{x}{\sigma}$	0	1	2	3	4	5	6	7	8	9
0.0	100000	99995	99980	99955	99920	99875	99820	99755	99685	99596
0.1	99501	99396	99283	99158	99025	98881	98728	98565	98393	98211
0.2	98020	97819	97609	97390	97161	96923	96676	96420	96156	95882
0.3	95600	95309	95010	94702	94387	94055	93723	93382	93024	92677
0.4	92312	91399	91558	91169	90774	90371	89961	89543	89119	88688
0.5	88250	87805	87353	86896	86432	85962	85488	85006	84519	84060
0.6	83527	83023	82514	82010	81481	80957	80429	79896	79359	78817
0.7	78270	77721	77167	76610	76048	75484	74916	74342	73769	73193
0.8	72615	72033	71448	70861	70272	69681	69087	68493	67896	67298
0.9	66689	66097	65494	64891	64287	63683	63077	62472	61865	61259
1.0	60653	60047	59440	58834	58228	57623	57017	56414	55810	55209
1.1	54607	54007	53409	52812	52214	51620	51027	50437	49848	49260
1.2	48675	48092	47511	46933	46357	45783	45212	44644	44078	43516
1.3	42956	42399	41845	41294	40747	40202	39661	39123	38569	38058
1.4	37531	37007	36487	35971	35459	34950	34445	33944	33447	32954
1.5	32465	31980	31500	31023	30550	30082	29618	29158	28702	28251
1.6	27804	27361	26923	26489	26059	25634	25213	24797	24385	23978
1.7	23575	23176	22782	22392	22008	21627	21251	20879	20511	20148
1.8	19790	19436	19086	18741	18400	18064	17732	17404	17081	16762
1.9	16448	16137	15831	15530	15232	14939	14650	14364	14083	13806
2.0	13534	13265	13000	12740	12483	12230	11981	11737	11496	11259
2.1	11025	10795	10570	10347	10129	09914	09702	09495	09290	09090
2.2	08892	08698	08507	08320	08136	07956	07778	07604	07433	07265
2.3	07100	06939	06780	06624	06471	06321	06174	06029	05888	05750
2.4	05614	05481	05350	05222	05096	04973	04852	04734	04618	04505
2.5	04394	04285	04179	04074	03972	03873	03775	03680	03586	03494
2.6	03405	03317	03232	03148	03066	02986	02908	02831	02757	02684
2.7	02612	02542	02474	02408	02343	02280	02218	02157	02098	02040
2.8	01984	01929	01876	01823	01772	01723	01674	01627	01581	01536
2.9	01492	01449	01408	01367	01328	01288	01252	01215	01179	01145
3.0	01111	00819	00598	00432	00309	00219	00153	00106	00073	00050
4.0	00034	00022	00015	00010	00006	00004	00003	00002	00001	00001
5.0	00000									

distribution by formula (20) is .244, and the distribution appears to be slightly leptokurtic (this is shown by the "peak" rising above the normal curve). Neither measure of divergence, however, is significant of a "real" discrepancy between our data and that of the normal distribution (see p. 220). On the whole, then, the normal curve plotted in Figure 26 fits the obtained distribution well enough to warrant our treating these data as sensibly normal.

IV. Why Frequency Distributions Deviate from the Normal Form

It is often important for the research worker to know why his distributions diverge from the normal form, and this is especially true when the deviation from normality is large and significant (p. 220). The reasons why distributions exhibit skewness and kurtosis are numerous and often complex, but a careful analysis of the data will often permit the setting up of hypotheses concerning the causes of non-normality which may be tested experimentally. The common causes of asymmetry, all of which must be taken into consideration by the careful experimenter, will be summarized in the present section.

1. Unrepresentative or Biased Sampling

Selection is a potent cause of asymmetry. We should hardly expect the distribution of I.Q.'s obtained from a group of twenty-five ten-year-old boys (all superior students) to be normal; nor would we look for symmetry in the distribution of I.Q.'s got from a special class of dull-normal ten-year-old boys, even though the group were fairly large. Neither of these groups is an unbiased selection (i.e., a cross-section) from the population of ten-year-old boys; and in addition, the first group is quite small. A small sample is not *necessarily* unrepresentative, but more often than not it is apt to be.

Selection will produce skewness and kurtosis in distributions even when the test has been adequately constructed and carefully administered. For example, a group of elementary school

pupils which contains (*a*) a large proportion of bilinguals, (*b*) many children of very low or very high socio-economic status, (*c*) a large number of pupils over-age for grade or accelerated, will almost surely return skewed distributions of test scores even upon standard intelligence and educational achievement examinations.

Scores made by small and homogeneous groups are likely to yield leptokurtic distributions; while scores from large and heterogeneous groups are more likely to be platykurtic. The distribution of scores achieved upon an educational examination by pupils throughout the elementary grades, as well as the distribution of chronological ages for these same pupils, will probably be somewhat flattened owing to the considerable overlap from grade to grade.

Distributions of physical traits, such as height, weight, and strength, are also affected by selection. Measurements of physical traits in large groups of the same age, sex, and race will closely approximate the normal form (p. 111). But the distribution of height for fourteen-year-old girls in the high school of a small city, or the distribution of weight for freshmen in a midwestern college will probably be skewed, as these groups are subject to selection in various traits related to height and weight.

2. Use of Unsuitable or Poorly Made Tests

If a test is too easy, scores will pile up at the high-score end of the distribution, while if the test is too hard scores will pile up at the low-score end. Imagine, for example, that an examination in arithmetic which requires only addition, subtraction, multiplication, and division, has been given to 1000 seventh graders. The resulting distribution will almost certainly be badly skewed to the left (see Figure 23). On the other hand, if the examination contains only problems in complex fractions, interest, square root, and the like, the score distribution is likely to be positively skewed — low scores will be more numerous than intermediate or high scores. It is probable also

that both distributions will be somewhat more "peaked" (leptokurtic) than the normal.

Asymmetry in cases like these may be explained in terms of those small positive and negative factors which determine the normal distribution. Too easy a test excludes from operation some of the factors which would make for an extension of the curve at the upper end, such as knowledge of more advanced arithmetical processes which the brighter child would know. Too hard a test excludes from operation factors which make for the extension of the distribution at the low end, such as knowledge of those very simple facts which would have permitted the answering of a few at least of the easier questions had these been included. In the first case we have a number of perfect scores and little discrimination; in the second case a number of zero scores and equally poor differentiation. Besides the matter of difficulty in the test, asymmetry may be brought about by ambiguous or poorly made items and by other technical faults.*

3. The Measurement of Traits the Distributions of Which Are Not Normal

Skewness or kurtosis or both may also appear owing to a real lack of normality in the trait being measured.† Non-normality of distribution will arise, for instance, when some of the hypothetical factors determining performance in a trait are dominant or prepotent over the others, and hence are present more often than chance will allow. Illustrations may be found in distributions resulting from the throwing of loaded dice. When off-center or biased dice are cast the resulting distribution will certainly be skewed and probably peaked,

* Hawkes, Lindquist and Mann, *The Construction and Use of Achievement Exams.* (1936), Chapters II and III.

† There is no reason why all distributions should approach the normal form. Thorndike has written: "There is nothing arbitrary or mysterious about variability which makes the so-called normal type of distribution a necessity, or any more rational than any other sort, or even more to be expected on *a priori* grounds. Nature does not abhor irregular distributions." — *Theory of Mental and Social Measurements* (1913), pp. 88–89.

owing to the greater likelihood of combinations of faces yielding extreme scores. The same is true of biased coins. Suppose, for example, that the probability of "success" (appearance of H) is four times the probability of failure (non-occurrence of H, or presence of T), so that $p = 4/5$, $q = 1/5$, and $(p + q) = 1.00$. If we think of the factors making for success or failure as 3 in number, we may expand $(p + q)^3$ to find the incidence of success and failure in varying degree. Thus, $(p + q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$, and substituting $p = 4/5$ and $q = 1/5$, we have

(1)

$$p^3 = (4/5)^3 = \frac{64}{125}$$

$$3p^2q = 3(4/5)^2 \cdot (1/5) = \frac{48}{125}$$

$$3pq^2 = 3(4/5) \cdot (1/5)^2 = \frac{12}{125}$$

$$q^3 = (1/5)^3 = \frac{1}{125}$$

(2) Expressed as a frequency distribution:

"Successes"	f
3	64
2	48
1	12
0	1
	$N = 125$

The numerators of the probability ratios (frequency of success) may be plotted in the form of a histogram to give Figure 27.

Note that this distribution is negatively skewed (to the *left*); that the incidence of three "successes" is 64, of two 48, of one 12, and of none 1. J-shaped distributions like these are essentially non-normal. Such curves have been most often found by psychologists to describe certain forms of social behavior. For example, suppose that we tabulate the number of students who appear at a lecture "on time"; and the number who come in five, ten, and fifteen-plus minutes late. If frequency of arrival is plotted against time, the distribution will be highest at zero ("on time") on the *Y-axis* and will fall off rapidly as we go to the right, i.e., will be positively skewed and J-shaped (see Figure 24). If only the early-comers are tallied, up to the "on time" group the curve will be a negatively skewed J-curve like those in Figures 23 and 27. J-curves describe behavior which is essentially non-normal in occurrence

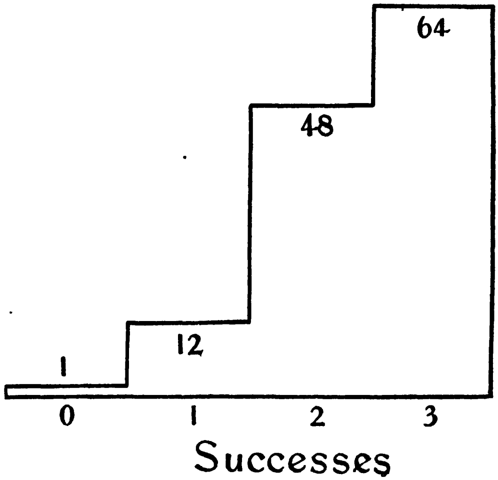


FIG. 27. Frequency Polygon of the expansion $(p + q)^3$, where $p = \frac{4}{5}$, $q = \frac{1}{5}$. p is the probability of success, q the probability of failure.

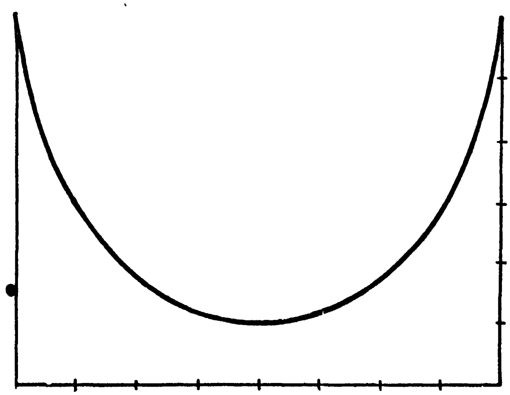


FIG. 28. U-shaped Frequency Curve.

because the causes of the behavior differ greatly in strength. But J-curves may also represent frequency distributions badly skewed for other reasons. We have seen in (1) and (2) above that selection and poorly chosen tests can produce distributions which closely resemble J-curves.

True J-curves often occur in medical statistics. The frequency of death due to degenerative disease, for instance, is highest during maturity and old age and minimal during the early years. If age is laid off on the baseline and frequency of death plotted on the *Y-axis* the curve will be negatively skewed and will resemble Figure 23 closely. Factors making for death are prepotent over those making for survival as age increases, and hence the curve is essentially asymmetrical. In the case of a childhood disease, the occurrence of death will be positively skewed when plotted against age as the probability of death becomes less with increase in age.

Another non-normal distribution, which may be mentioned briefly, is the U-shaped curve shown in Figure 28. U-shaped distributions, like J-curves, are probably more often encountered in the measurement of social and personality traits than in the measurement of mental abilities. Suppose, for instance, that the distribution of a large group of college freshmen upon an intelligence examination has been drawn up. Now, if the proportion in each score category who report more than a stipulated number of "neurotic" symptoms is determined, it is likely that the high- and low-scoring students will report more symptoms than the intermediate-scoring students. Accordingly, the curve for symptoms will be U-shaped, will rise at both ends. Again, suppose that all pupils in an elementary school *below* I.Q. 75 and *above* I.Q. 120 are taught in special classes. Then, since the *total* number of such children will probably be largest in the low and high grades, a plot of pupils by grades will tend to be *U*-shaped.

4. The Influence upon Distribution Form of Errors Made in the Construction and Administration of Tests

There are a number of factors besides those already mentioned which make for asymmetry in score distributions. Differences in the size of the units in which a trait has been measured, for example, will lead to skewness. Thus, if the test items are very easy at the beginning and very hard later on, an increment

of one point of score at the upper end of the test scale will be much greater than an increment of one point at the low end of the scale. The effect of such unequal or "rubbery" units is the same as that encountered when the test is too easy — scores tend to pile up toward the high end of the scale and be stretched out or skewed toward the low end.

Errors in administration of a test as in timing or giving instructions; errors in the use of scoring stencils; large differences in practice or in motivation among the subjects — all of these factors, if they cause many students to score higher or lower than they normally would, will make for asymmetry in the distribution.

PROBLEMS

1. In two throws of a coin, what is the probability of throwing at least one head?
2. What is the probability of throwing exactly one head in three throws of a coin?
3. Five coins are thrown. What is the probability that exactly two of them will be heads?
4. If the probability of answering a certain question correctly is four times the probability of answering it incorrectly, what is the probability of answering it correctly?
5. A rat has five choices to make of alternate routes in order to reach the food-box. If it is true that for each choice the odds are two to one in favor of the correct pathway, what is the probability that the rat will make all of its choices correctly?
6. Assume that trait X is completely determined by 6 factors — all similar and independent, and each as likely to be present as absent — plot the distribution which one might expect to get from the measurement of trait X in an unselected group of 1000 people.
7. Toss five pennies thirty-two times, and record the number of heads and tails after each throw. Plot frequency polygons of obtained and expected occurrences on the same axes. Compare the M's and σ's of obtained and expected distributions.

8. What percentage of a normal distribution is included between the
   (a) mean and 1.54σ
   (b) mean and $-2.7PE$
   (c) -1.73σ and $.56\sigma$
   (d) $-3.5PE$ and $1.0PE$
   (e) $.66\sigma$ and 1.78σ
   (f) $-1.8PE$ and $-2.5PE$

9. In a normal distribution
   (a) Determine P_{27}, P_{46}, P_{54}, and P_{81} in σ-units.
   (b) What are the percentile ranks of scores at -1.23σ, $-.50\sigma$, $+.84\sigma$?

10. (a) Compute measures of skewness and kurtosis for each of the four frequency distributions in Chapter II, Problem 1, page 46.
    (b) Fit normal probability curves to these same distributions, using the method given on page 123.
    (c) For each distribution, compare the percentage of cases lying between $\pm 1\sigma$ with the 68.26% found in the normal distribution.

Answers

1. 3/4 2. 3/8 3. 10/32 4. 4/5 5. 32/243

7. For expected distribution
   $M = 2.5$, $\sigma = 1.12$

8. (a) .4383 (d) .7409
   (b) .4657 (e) .2171
   (c) .6705 (f) .0665

9. (a) $-.61\sigma$, $-.10\sigma$, $.10\sigma$, $.88\sigma$
   (b) 11, 31, 80

10. (a)

	Skewness		*Kurtosis*
	By formula (18)	By formula (19)	By formula (20)
(1)	− .018	− .27	.239
(2)	.156	1.03	.277
(3)	.071	.55	.222
(4)	.032	− .35	.248

(c) 66%, 67%, 66%, 66%

CHAPTER VI

APPLICATIONS OF THE NORMAL PROBABILITY CURVE

I. Problems Involving Proportions of Area within Different Parts of the Normal Distribution

This section will consider a number of problems which may be readily solved if we can assume that the distributions of scores with which we are dealing may be treated as normal, or at least as approximately normal, in form. Each general problem will be illustrated by several examples. These examples are intended to present the issues concretely, and should be carefully worked through by the student. Constant reference will be made to Tables 17 and 18; and a knowledge of how to use these tables is essential.

1. To Determine the Percentage of Cases in a Normal Distribution Which Fall within Given Limits

Example (*1*) Given a normal distribution with a mean of 12, and a σ of 4. (*a*) What percentage of the cases fall between 8 and 16? (*b*) What percentage of the cases lie above 18? (*c*) Below 6?

(*a*) A score of 16* is four points above the mean, and a score of 8 is four points below the mean. If we divide this scale distance of four score units by the σ of the distribution (i.e., by 4) it is clear that 16 is 1σ above the mean, and that 8 is 1σ below the mean (see Fig. 29, p. 136). There are 68.26% of the cases in a normal distribution between the mean and $\pm 1\sigma$ (Table 17). Hence, 68.26% of the scores in this distribution, or approximately the middle two-thirds, fall between 8 and 16.

* A score of 16 is the midpoint of the interval 15.5 to 16.5.

This result may also be stated in terms of "chances." Since 68.26% of the cases in the given distribution fall between 8 and 16, the chances are about 68 in 100 that any score in the distribution will be found between these limits.

(*b*) A score of 18 is six score units, or 1.5σ above the mean $(6/4 = 1.5)$. From Table 17 we find that 43.32% of the cases in the entire distribution fall between the mean and 1.5σ. Accordingly, 6.68% of the cases (.5000–.4332) must lie above 18, in order to fill out the 50% of cases in the upper half of the curve

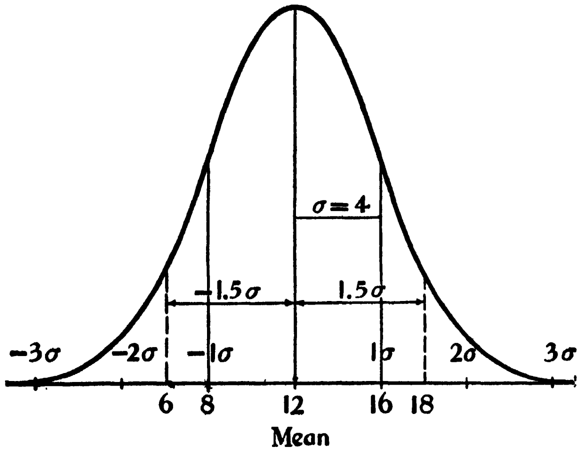


FIG. 29.

(Fig. 29). Stated in terms of chances, there are 668 chances in 10,000, or about 7 in 100, that any score in the distribution will lie above 18.

(*c*) A score of 6 is -1.5σ from the mean. Between the mean and a score of $6(-1.5\sigma)$ are 43.32% of the cases in the whole distribution. Hence, about 7% of the cases lie below 6 (fill out the 50% below the mean), and the chances are 7 in 100 that any score in the distribution will fall below 6.

Example (*2*) Given a normal distribution with a mean of 29.75, and a Q of 4.56. What percentage of the distribution lie between 22 and 26? What are the chances that a score will fall between 22 and 26?

In a normal distribution $Q = PE$. A score of 22 is 7.75 units, or $-1.70PE$ ($7.75/4.56 = 1.70$) from the mean; and a score of 26 is 3.75 or $-.82PE$ from the mean (Fig. 30, below). From Table 18 we know that 37.42% of the cases in a normal distribution lie between the mean and $-1.70PE$; and that 20.99% (by interpolation) of the cases lie between the mean and $-.82PE$. By simple subtraction, therefore, 16.43% of the cases fall between $-1.70PE$ and $-.82PE$ or between 22 and 26. The chances are about 16 in 100 that a score will fall between 22 and 26.

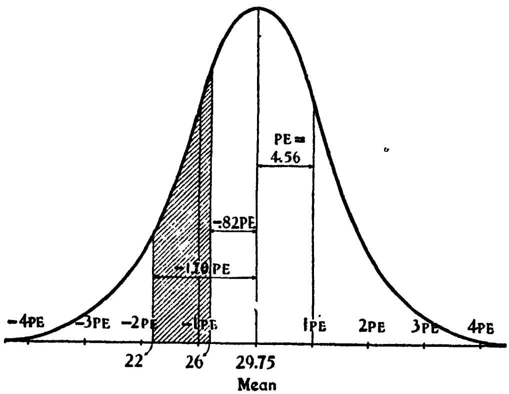


FIG. 30.

2. To Find the Limits in Any Normal Distribution Which Will Include a Given Percentage of the Cases

Example (*1*) Given a normal distribution with a mean of 16.00 and a σ of 4.00. What limits will include the middle 75% of the cases?

The middle 75% of the cases in a normal distribution must include the 37.5% just above, and the 37.5% just below the mean. From Table 17 we find that 3749 cases in 10,000, or 37.5% of the distribution, fall between the mean and 1.15σ; and, of course, 37.5% of the distribution also fall between the mean and -1.15σ. The middle 75% of the cases, therefore,

lie between the mean and $\pm$ 1.15σ; or, since σ = 4.00, between the mean and $\pm$ 4.60 score units. Adding $\pm$ 4.60 to the mean (to 16.00), we find that the middle 75% of the scores in the given distribution lie between 20.60 and 11.40 (see Fig. 31, below).

Example (2) Given a normal distribution with a median of 150.00 and a Q of 26.00. What limits will include the *highest* 20% of the distribution? The *lowest* 10%?

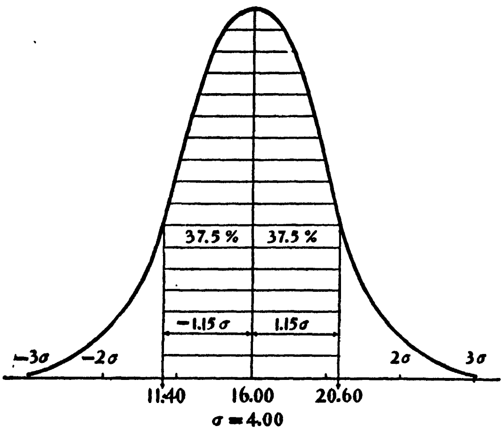


FIG. 31.

The highest 20% of a normally distributed group will have 30% of the cases between its lower limit and the median, since 50% of the cases lie in the right half of the distribution. From Table 18, we know that 3004 cases in 10,000, or 30% of the distribution, fall between the median and 1.25PE. Since the PE of the given distribution is 26.00, 1.25PE will be 1.25 × 26.00 or 32.5 points *above* the median, namely, at 182.5. The lower limit of the highest 20% of the given group, therefore, is 182.5; and the upper limit is the highest score in the distribution, whatever that may be.

The lowest 10% of a normally distributed group will have 40% of the cases between the median and its upper limit. Exactly 40% of the distribution fall between the median and − 1.90PE. The PE of the given distribution is 26.00; hence,

$-1.90PE$ will be 1.90×26.00 or 49.4 score units *below* the median, that is at 100.6. The upper limit of the lowest 10% of scores in the given group, therefore, is 100.6; and the lower limit is the lowest score in the distribution.

3. To Compare Two Distributions in Terms of "Overlapping"

- *Example* (*1*) Given the distributions of the scores made on a logical memory test by 300 boys and 250 girls (Table 21). The boys' mean score is 21.49 with a σ of 3.63. The girls' mean score is 23.68 with a σ of 5.12. The medians are: boys, 21.41, and girls, 23.66. What percentage of boys exceed the median of the girls' distribution?

On the assumption that these distributions are sensibly normal, we may solve this problem by means of Table 17. The girls' *median* is $23.66 - 21.49$ or 2.17 score units above the boys' *mean*. Dividing 2.17 by 3.63 (the σ of the boys' distribution), we find that the girls' median is $.60\sigma$ above the mean of the boys' distribution. Table 17 shows that 23% of a normal distribution lie between the mean and $.60\sigma$; hence 27% of the boys (50% − 23%) exceed the girls' median.

This problem may also be solved by direct calculation from the distributions of boys' and girls' scores without any assumption as to normality of distribution. The calculations are shown in Table 21; and it will be interesting to compare the result found by direct calculation with that obtained by use of the probability tables. The problem is to find the *number* of boys whose scores exceed 23.66, the girls' median, and then turn this number into a percentage. There are 217 boys who score up to 23.5 (lower limit of 23.5 to 27.5). The class-interval 23.5 to 27.5 contains 68 scores; hence there are 68/4 or 17 scores *per scale unit* on this interval. We wish to reach 23.66 in the boys' distribution. This point is .16 of a score ($23.66 - 23.50 = .16$) above 23.5, or 2.72 (i.e., $17 \times .16$) score units above 23.5. Adding 2.72 to 217, we find that 219.72 of the boys' scores fall *below* 23.66, the girls' median. Since $300 - 219.72 = 80.28$, it is clear that $80.28 \div 300$ or 26.76% (approximately 27%) of the

TABLE 21

TO ILLUSTRATE THE METHOD OF DETERMINING OVERLAPPING BY DIRECT CALCULATION FROM THE DISTRIBUTION

Boys Scores	f	*Girls* Scores	f
27.5 to 31.5	15	31.5 to 35.5	20
23.5 to 27.5	68	27.5 to 31.5	35
19.5 to 23.5	128	23.5 to 27.5	73
15.5 to 19.5	79	19.5 to 23.5	68
11.5 to 15.5	10	15.5 to 19.5	41
	$N = 300$	11.5 to 15.5	13
	$N/2 = 150$		$N = 250$
			$N/2 = 125$

$$Mdn = 19.5 + \tfrac{61}{128} \times 4 = 21.41 \qquad M = 21.49 \qquad \sigma = 3.63$$

$$Mdn = 23.5 + \tfrac{3}{73} \times 4 = 23.66 \qquad M = 23.68 \qquad \sigma = 5.12$$

What percent of the boys exceed 23.66, the median of the girls? First, 217 boys make scores *below* 23.5. The class-interval 23.5–27.5 contains 68 scores; hence, there are 68/4 or 17 scores *per scale unit* on this interval.

The girls' median, 23.66, is .16 *above* 23.5, lower limit of interval 23.5–27.5. If we multiply 17 (number of scores per scale unit) by .16 we obtain 2.72 which is the distance we must go into interval 23.5–27.5 to reach 23.66.

Adding 217 and 2.72, we obtain 219.72 as that part of the boys' distribution which falls *below* the point 23.66 (girls' median). N is 300; hence 300–219.72 gives 80.28 as that part of the boys' distribution which lies *above* 23.66. Dividing 80.28 by 300, we find that .2676, or approximately 27%, of the boys exceed the girls' median.

boys exceed the girls' median. This result is in almost perfect agreement with that obtained above. Apparently the assumption of normality of distribution for the boys' scores was justified.

The agreement between the percentage of overlapping found by direct calculation from the distribution, and that found by use of the probability tables will nearly always be close, especially if the groups are large and the distributions fairly symmetrical. When the overlapping distributions are small and not very regular in outline, it is safer to use the method of direct calculation since no assumption as to form of distribution is then made.

4. To Determine the Relative Difficulty of Test Questions, Problems, and other Test Items

Example (1) Given a test question or problem solved by 10% of a large unselected group; a second problem solved by 20% of the same group; and a third problem solved by 30%. If we assume the capacity measured by the test problems to be distributed normally, what is the relative difficulty of questions 1, 2, and 3?

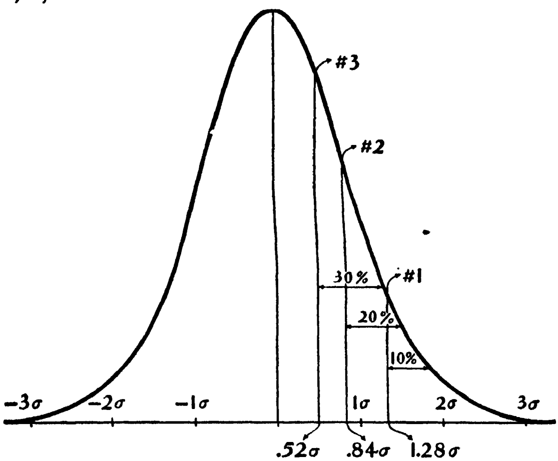


FIG. 32.

Our first task is to find for Question 1 a position in the distribution, such that 10% of the entire group (the percent passing) lie above, and 90% (the percent failing) lie below the given point. The highest 10% in a normally distributed group has 40% of the cases between its lower limit and the mean (see Fig. 32, above). From Table 17 we find that 39.97% (i.e., 40%) of a normal distribution fall between the mean and 1.28σ. Hence, Question 1 belongs at a point on the baseline of the curve, a distance of 1.28σ from the mean; and, accordingly, 1.28σ may be set down as the difficulty value of this question.

Question 2, passed by 20% of the group, falls at a point in the distribution 30% above the mean. From Table 17 it is

found that 29.95% (i.e., 30%) of the group fall between the mean and $.84\sigma$; hence, Question 2 has a difficulty value of $.84\sigma$. Question 3, which lies at a point in the distribution 20% above the mean, has a difficulty value of $.52\sigma$, since 19.85% of the distribution fall between the mean and $.52\sigma$. To summarize our results:

Question	Passed by	σ-value	σ-difference
1	10%	1.28	—
2	20%	.84	.44
3	30%	.52	.32

The σ-difference in difficulty between Questions 2 and 3 is .32, which is roughly 3/4 of the σ-difference in difficulty between Questions 1 and 2. Since the percentage difference is the same in the two comparisons, it is evident that when ability is assumed to follow the normal distribution, σ and not percentage differences are the better indices of differences in difficulty.

Example (2) Given three test items, 1, 2, and 3, passed by 50%, 40%, and 30%, respectively, of a large group. On the assumption of normality of distribution, what percentage of this group must pass test item 4, in order for it to be as much more difficult than 3, as 2 is more difficult than 1?

An item passed by 50% of a group is, of course, failed by 50%; and, accordingly, such an item falls exactly in the middle of a normal distribution of "difficulty." Test item 1, therefore, has a σ-value of .00 since it falls exactly at the mean (Fig. 33). Test item 2 lies at a point in the distribution 10% above the mean, since 40% of the group passed, and 60% failed this item. Accordingly, the σ-value of item 2 is .25, since from Table 17 we find that 9.87% (roughly 10%) of the cases lie between the mean and $.25\sigma$. Test item 3, passed by 30% of the group, lies at a point 20% above the mean, and this item has a difficulty value of $.52\sigma$, as 19.85% (20%) of the normal distribution fall between the mean and $.52\sigma$.

Since item 2 is $.25\sigma$ farther along on the difficulty scale (to-

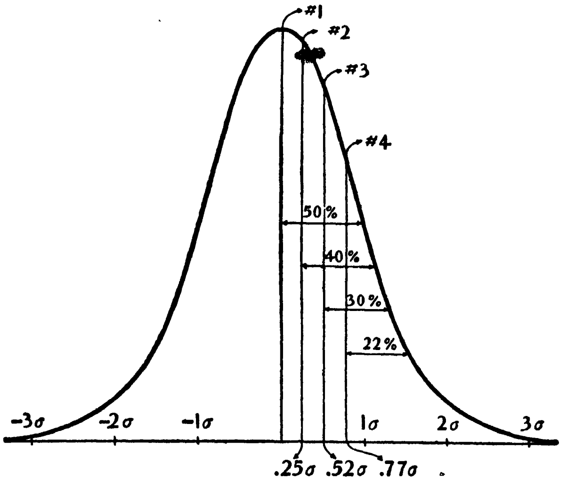


FIG. 33.

ward the high-score end of the curve) than item 1, it is clear that item 4 must be $.25\sigma$ above item 3, if it is to be as much harder than item 3 as item 2 is harder than item 1. Item 4, therefore, must have a value of $.52\sigma + .25\sigma$ or $.77\sigma$; and from Table 17 we find that 27.94% (28%) of the distribution fall between the mean and this point. This means that 50% − 28% or 22% of the group must *pass* item 4. To summarize:

Test Item	Passed by	σ-value	σ-difference
1	50%	.00	—
2	40%	.25	.25
3	30%	.52	—
4	22%	.77	.25

A test item, therefore, must be passed by 22% of the group in order for it to be as much more difficult than an item passed by 30%, as an item passed by 40% is more difficult than one passed by 50%. Note again that percentage differences are not reliable indices of differences in difficulty when the capacity measured is distributed normally.

5. To Separate a Given Group into Sub-Groups According to Capacity, When the Trait Measured Is Assumed to be Normally Distributed

Example (*1*) Suppose that we have administered a certain examination to 100 college students. We wish to classify our group into five sub-groups A, B, C, D, and E according to ability, the *range* of ability to be equal in each sub-group. On the assumption that the trait measured by our examination is normally distributed, how many students should be placed in groups A, B, C, D, and E?

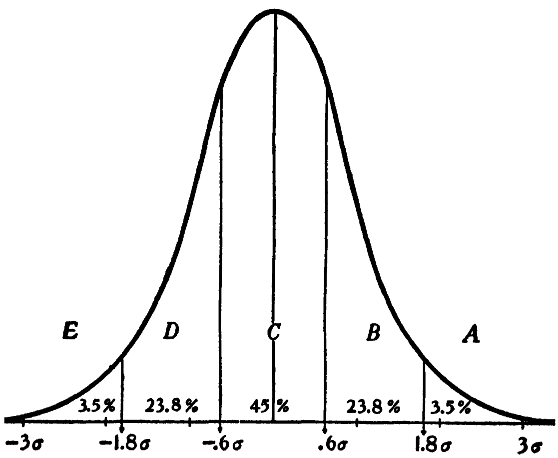


FIG. 34.

Let us first represent the positions of the five sub-groups diagrammatically on a normal curve as shown in Figure 34, above. If the baseline of the curve is considered to extend from -3σ to $+3\sigma$, that is, over a range of 6σ, dividing this range by 5 (the number of sub-groups) gives 1.2σ as the baseline extent to be allotted to each group. These five intervals may be laid off on the baseline as shown in the figure, and perpendiculars erected to demarcate the various sub-groups. Group A covers the upper 1.2σ; group B the next 1.2σ; group C lies $.6\sigma$ to the *right* and $.6\sigma$ to the *left* of the mean; groups D and E occupy

the same relative positions in the lower half of the curve that B and A occupy in the upper half.

To find what percentage of the whole group belongs in A we must find what percentage of a normal distribution lies between 3σ (upper limit of the A group) and 1.8σ (lower limit of the A group). From Table 17 49.86% of a normal distribution is found to lie between the mean and 3σ; and 46.41% between the mean and 1.8σ. Hence, 3.5% of the total area under the normal curve (49.86% − 46.41%) lie between 3σ and 1.8σ; and, accordingly, group A comprises 3.5% of the whole group.

The percentages in the other groups are calculated in the same way. Thus, 46.41% of the normal distribution fall between the mean and 1.8σ (upper limit of group B) and 22.57% fall between the mean and $.6\sigma$ (lower limit of group B). Subtracting, we find that 46.41% − 22.57% or 23.84% of our distribution belongs in sub-group B. Group C lies from $.6\sigma$ above to $-.6\sigma$ below the mean. Between the mean and $.6\sigma$ are 22.57% of the normal distribution, and the same percent lies between the mean and $-.6\sigma$. Group C, therefore, includes 45.14% (22.57 × 2) of the distribution. Finally, sub-group D which lies between $-.6\sigma$ and -1.8σ contains exactly the same percentage of the distribution as sub-group B; and group E, which lies between -1.8σ and -3σ, contains the same percent of the whole distribution as group A. The percentage and number of men in each group are given in the following table:

	Groups				
	A	B	C	D	E
Percent of total in each group	3.5	23.8	45	23.8	3.5
Number in each group (100 men in all)	4 or 3	24	45	24	4 or 3

On the assumption that the capacity measured follows the normal curve, it is clear that three to four men in our group of 100 should be placed in group A, the "marked" ability group; twenty-four in group B, the "high average" ability group; forty-five in group C, the "average" ability group; twenty-four

in group D, the "low average" ability group; and three or four in group E, the "very low" or "inferior" group.

The above procedure may be used to determine how many students in a class should be assigned to each of any given number of grade-groups. It must be remembered that the assumption is made that performance in the subject matter upon which the individuals are being marked is represented by the normal curve. The larger and more unselected the group the more nearly is this assumption justified.

II. The Scaling of Test Items

1. The Arrangement of Test Items into a Scale in Which the Difficulty of Each Item Is Known with Reference to an Arbitrary Zero Point

The psychologist often wishes to construct scales which shall contain problems or questions graded in difficulty from very easy to very difficult by known steps or intervals. Given a set of problems or test items, if we know what proportion of a large group passes each problem it is comparatively easy to arrange the problems in a percentage order of difficulty. Such an arrangement constitutes a "scale," to be sure; but it is a very crude scale, since we know only roughly the steps in difficulty from item to item.

In constructing scaled tests, the σ or *PE* of the distribution, rather than the percent passing, is taken as the unit of measurement. When the variability of the group is employed as a scaling unit, we are able not only to arrange test items in order of difficulty but to "set" or space them at definite points along a difficulty scale. To illustrate how test items are scaled when the unit of measurement is the σ or *PE* of the group, let us suppose that we wish to construct a scale for measuring "reasoning ability" (e.g., by means of syllogisms) in twelve-year-old children; or a test of arithmetic problems for Grade IV; or a scale for testing sentence memory in eight-year-old children. The successive steps involved in constructing such a scale may be outlined as follows:

(1) First compile a large number of problems or other test items. These items should vary in difficulty from very easy to very hard and should be representative of the field covered by the test.

(2) Administer the items or problems to as large and as randomly selected a group as can be assembled from among those for whom the test is eventually intended.

(3) Compute the percentage of the group solving each problem correctly. Duplicate items and those too easy or too hard or unsatisfactory for one reason or another should be discarded. The problems retained for the scale are then arranged in order of percentage difficulty. A problem solved correctly by 90% of the group is obviously less difficult than one solved correctly by 75%; while the second problem is, in turn, clearly less difficult than one solved correctly by 50%. The greater the percentage passing an item, the lower the position of this item in a scale of difficulty.

(4) By means of Table 18 convert the percentage solving each problem correctly into *PE* distances above or below the mean.* The procedure in detail is as follows: A problem solved correctly by 40% of the group is 10% or about .40*PE* above the mean. A problem solved correctly by 78% of the group is 28% (78% − 50%) or 1.15*PE* below the mean. We may tabulate the results for five items, selected at random, as follows (see Fig. 35, below):

Problems	A	B	C	D	E
Percent solving	93	78	55	40	14
Distance from mean in percentage terms	− 43	− 28	− 5	10	36
Distance from mean in *PE* terms	− 2.20	− 1.15	− .20	.40	1.60

Problem A is solved by 93% of the group, i.e., by the upper 50% (the right half of the curve) plus the 43% to the left of the mean. This puts problem A at a point − 2.20*PE*

* The procedure is identical when σ is employed instead of *PE*.

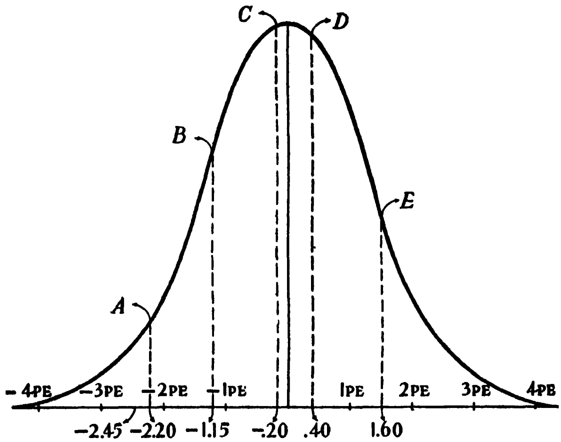


FIG. 35.

from the mean. In the same way, the percentage distance of each problem from the mean (measured in the plus or minus direction) is found by subtracting the percentage passing from 50%. From these percentages, the *PE* distance of the problem above or below the mean can be read from Table 18.*

(5) When the *PE* distance of each problem above or below the mean has been established, calculate the *PE* distance of each problem from the "zero point" of ability in the test. A zero point may be located in the following way: Suppose that 5% of the whole group fails to solve a single problem correctly. This would put the level of zero ability in this test 45% of the distribution below the mean, or at a point $-2.45PE$ from the mean.† The *PE* distance of each problem in the scale may now be calculated from this

* *PE*'s are taken for the percentage nearest to the given value, without interpolation.

† This value is an arbitrary, not a true, zero. It serves, however, as a convenient reference point (point of minimum ability) from which to measure performance. The points $-4.00PE$ or -3.00σ are also convenient reference points.

arbitrary zero point. To illustrate with the five problems above:

Problems	A	B	C	D	E
PE distance from mean........	– 2.20	– 1.15	– .20	.40	1.60
PE distance from arbitrary zero, i.e. – 2.45*PE*	.25	1.30	2.25	2.85	4.05

The simplest method of finding *PE* distances from a given zero is to subtract the zero point algebraically from the *PE* distance of each problem from the mean. Problem A, for example, is – 2.20 – (– 2.45) or .25*PE* from the arbitrary zero point; and Problem E is 1.60 – (– 2.45) or 4.05*PE* from the zero point. The *PE* value of each of the other problems, as measured from the arbitrary zero point, is found in the same way. When the *PE* value-from-zero of each of the problems intended for the test has been determined, the difficulty value of each problem with respect to every other problem, as well as with respect to the arbitrary zero, is known, and the scale is finished.

2. Scaling Total Scores on a Test

(1) Normalizing a Frequency Distribution: The *T*-Scale

In the last section we saw how separate test items are scaled in *PE*-units on the assumption of normality in the trait measured. We shall now describe a method of scaling score totals or aggregates of items — a procedure usually followed in standard educational achievement tests.

The method consists essentially in "normalizing" the distribution of test scores. This is done by transforming original test scores into equivalent scores in a normal distribution. Equivalent scores are defined as measures which indicate the same levels of ability. Suppose that in a given test 84% of the group score below 124. Then 124 is equivalent to a score of $+ 1\sigma$ in a normal distribution, since 84% (approximately) of a normal distribution fall below (to the left of) $+ 1\sigma$. As we shall see later, normalizing a distribution of test scores alters

the original test units (stretching them out or compressing them) and the more skewed the original distribution ~~the greater the~~ change in unit.

The obtained scores of a distribution may be transformed into various systems of "new" or normalized scores. The method outlined in this section leads to a normalized system of scores called *T*-scores. *T*-scaling was devised by McCall* and first used by him in constructing a series of reading tests designed for use in the elementary grades. The original *T*-scale was drawn up from the scores achieved by 500 twelve-year-olds upon a reading test; and the scores made by other age groups on these tests were expressed in terms of twelve-year-old *T*-scores. Since the first use of the method, *T*-scaling has been employed with various groups and no longer has reference specifically to twelve-year-olds or to reading tests.

Procedure in *T*-scaling can best be shown by an example. We shall outline the process in a series of steps and illustrate each step by reference to the data in Table 22.

TABLE 22

To Illustrate the Calculation of T-Scores

(1) Test Score	(2) f	(3) Cum. f	(4) Cum. Freq. below Score + $\frac{1}{2}$ on Given Score	(5) Col. (4) in %'s	(6) *T*-Scores
10	1	62	61.5	99.2	74
9	4	61	59	95.2	67
8	6	57	54	87.1	61
7	10	51	46	74.2	56
6	8	41	37	59.7	52
5	13	33	26.5	42.7	48
4	18	20	11	17.7	41
3	2	2	1	1.6	28
	$N = 62$				

(1) Compile a large and representative group of test items which vary in difficulty from easy to hard. Administer these

* McCall, W. A., *How to Measure in Education* (1929), Chapter X, pp. 272–306.

items to a sample of subjects (children or adults) for whom the test is intended eventually.

(2) Compute the percent passing each item. These percents may be converted into σ-units so that the items selected for inclusion in the final test are arranged in order of difficulty in terms of σ. Since a precise measure of relative difficulty is not important at this stage, however, items in the final test may be arranged simply in order of percentage difficulty (number passing).

(3) Administer the final test to a representative sample and tabulate the distribution of scores. Total scores may be scaled as shown in Table 22 for a group of sixty-two subjects. In column (1) of Table 22 the test scores are entered. In column (2) are the frequencies, i.e., numbers of subjects who achieve various scores. Two subjects, for example, had scores of 3, 18 scores of 4, 13 scores of 5, and so on. In column (3) scores have been cumulated (p. 74) from the low to the high end of the frequency distribution. Column (4) shows the number of subjects who fall *below* each score plus one half of those who achieve the given score. The entries in this column may be computed readily from columns (2) and (3). Since there are no scores below 3 and two scores on 3, the number below plus one-half on 3, is 1. There are two scores below 4 [column (3)] and eighteen on 4 [column (2)], hence the number below plus one-half on 4 is 9 + 2, or 11. There are twenty scores below 5 [column (3)], and thirteen scores on 5 [column (2)], hence the number below plus one-half on 5 is 20 + 6.5, or 26.5. One-half of the frequency *on* a given score must be added to the number of scores falling *below* the score because a score is an interval, not a point. The score of 4, for example, is the interval 3.5 to 4.5, mid-point 4.0. If the eighteen frequencies on 4 are thought of as distributed evenly over the interval, nine will lie *below* and nine *above* 4.0, the midpoint. Hence, if we add nine to the two scores below 4 (i.e., below 3.5), we obtain eleven as the number of scores *below* 4.0, the midpoint of the

interval 3.5 to 4.5. Each sum in column (4) is up to the mid-point of a score-interval.

In column (5) the entries in column (4) are expressed as percentages of N (62). Thus 99.2% of the scores lie below 10.0, midpoint of 9.5 to 10.5; 95.2% of the scores lie below 9.0, etc.

(4) Turn the percents in column (5) into T-scores by means of Table 23. T-scores (to two places) in Table 23 corresponding to percentages nearest those wanted are taken without interpolation, as fractional T-scores are a needless refinement. Thus 1.39 (T-score = 28) is taken for 1.6; 18.41 (T-score = 41) for 17.7, and so on.

Figure 36 shows a histogram plotted from the distribution of the sixty-two scores in Table 22. Note that the scores 3, 4,

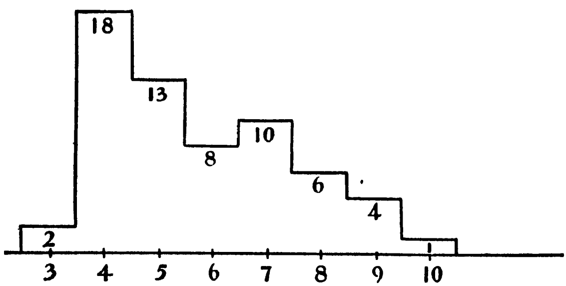


FIG. 36. Histogram of the Sixty-two Scores in Table 22.

5, etc., are spaced at equal intervals along the baseline, i.e., along the scale of scores. When these scores are transformed into equivalent normal curve scores — into T-scores — they occupy the positions in the normal curve shown in Figure 37. The unequal scale distances between the scores in Figure 37 show clearly that, on the assumption of normality in the trait, the original scores do not represent equal difficulty steps.

T-scores are simply σ-scores in a normal distribution multiplied by 10 and referred to an arbitrary reference point below the mean in order to avoid negative signs. In the σ scaling of items, the mean is taken at zero and σ is put equal to 1. The

TABLE 23

To Facilitate the Calculation of T-Scores

The percents refer to the percentage of the total frequency below a given score + 1/2 of the frequency on that score. *T*-scores are read directly from the given percentages.

Percent	*T-score*	*Percent*	*T-score*
.0032	10	53.98	51
.0048	11	57.93	52
.007	12	61.79	53
.011	13	65.54	54
.016	14	69.15	55
.023	15	72.57	56
.034	16	75.80	57
.048	17	78.81	58
.069	18	81.59	59
.097	19	84.13	60
.13	20	86.43	61
.19	21	88.49	62
.26	22	90.32	63
.35	23	91.92	64
.47	24	93.32	65
.62	25	94.52	66
.82	26	95.54	67
1.07	27	96.41	68
1.39	28	97.13	69
1.79	29	97.72	70
2.28	30	98.21	71
2.87	31	98.61	72
3.59	32	98.93	73
4.46	33	99.18	74
5.48	34	99.38	75
6.68	35	99.53	76
8.08	36	99.65	77
9.68	37	99.74	78
11.51	38	99.81	79
13.57	39	99.865	80
15.87	40	99.903	81
18.41	41	99.931	82
21.19	42	99.952	83
24.20	43	99.966	84
27.43	44	99.977	85
30.85	45	99.984	86
34.46	46	99.9890	87
38.21	47	99.9928	88
42.07	48	99.9952	89
46.02	49	99.9968	90
50.00	50		

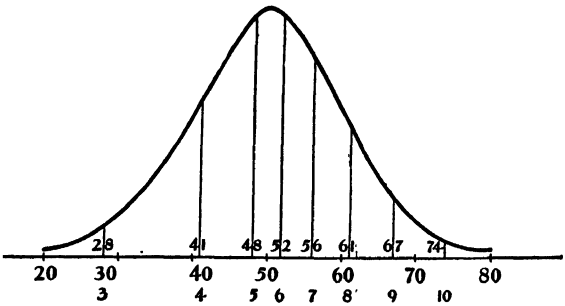


FIG. 37. Normalized Distribution of the Scores in Table 22 and Figure 36. Original scores and T-score equivalents are shown on baseline.

point of reference, therefore, is zero and the unit of measurement is one. Now if the point of reference is moved from the mean of the normal curve to a point -5σ below the mean, this new reference point becomes zero and the mean becomes five. The σ divisions above the mean ($+1\sigma$, $+2\sigma$, $+3\sigma$, $+4\sigma$, and $+5\sigma$) become 6, 7, 8, 9, and 10; and the σ divisions below the mean (-1σ, -2σ, -3σ, -4σ, and -5σ) are 4, 3, 2, 1, and 0. The σ of the distribution remains, of course, equal to 1, as shown in Figure 38.

Relatively slight changes are needed in order to convert this σ-scale into a T-scale. The T-scale begins at -5σ and ends at $+5\sigma$. But σ is multiplied by 10 so that the mean is 50 and the other σ divisions are 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. The relationship of the T-scale to the ordinary σ-scale is shown in Figure 38. Note that the T-scale ranges from 0 to 100; that its unit (T) is 1 (i.e., .1 of σ which is taken equal to 10), and that the mean is 50. The reference point on the T-scale is set at -5σ in order to have the scale cover exactly 100 units. This is convenient, but it puts the extremes of the scale far beyond the ability ranges of most groups. In actual practice, T-scores range from about 15 to 85 (i.e., from -3.5σ to $+3.5\sigma$).

In Table 23, percents lying to the *left* of (below) succeeding σ-points expressed as T-scores are tabulated, rather than percents between the mean and given σ-points as in Table 17.

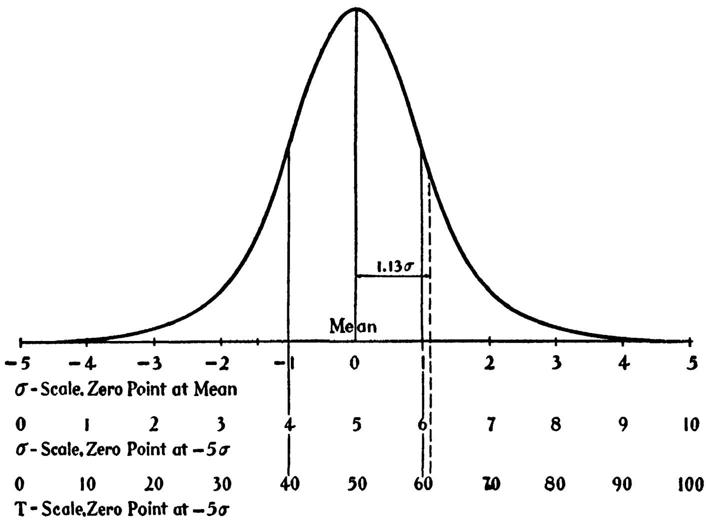


FIG. 38. To Illustrate σ-Scaling and T-Scaling in a Normal Distribution.

Table 23 is useful, therefore, in enabling one to read T-scores directly, but the reader should note that T-scores can also be computed from Table 17. We may illustrate with the score of 8 (Table 22) which has a percent below plus one-half reaching of 87.1. A score *failed* by 87.1% lies (87.1 − 50) or 37.1% to the *right* of the mean. From Table 17 we read that 37.1% of the distribution lies between the mean and 1.13σ. Since the σ of the T-scale is 10, 1.13σ becomes 11 in T-units; and adding 11 to 50, the mean, we obtain 61 as the required T-score (see Fig. 38).

T-scores are expressed in terms of the same unit and with respect to the same reference point; and unlike percentiles are equal over the scale. T-scaling is superior to the method of scaling separate items because the difficulty value of a score is more stable than the difficulty value of a single item. T-scales, too, have the advantage that scores ranging from 0 to 100 are more readily understood than are σ-scores expressed in other units.

(2) A Comparison of *T*-scores and Standard (*Z*) Scores

T-scores are sometimes confused with standard scores, but the assumptions underlying the computation of the two sorts of measures are quite different. Table 24 repeats the original data of Table 22 and shows the *T*-score equivalents to given "raw" scores. Standard scores, denoted by *Z*, are listed in column (4) for comparison with the *T*-scores. These *Z* scores were calculated in the following way: The mean of the original distribution is 5.73 and the σ is 1.72. Each score in the distribution may be expressed as a σ-deviation from the mean. Thus $\frac{3 - 5.73}{1.72} = \frac{-2.73}{1.72} = -1.59\sigma$; $\frac{4 - 5.73}{1.72} = -1.00\sigma$, and so on for the others. These σ-scores may be transformed into a new distribution with any mean and σ we wish. Suppose we set the "new" mean at 50 and the "new" σ at 10 (as in the *T*-scale). Then the score of 3 is -1.59×10 or -16 units from 50 or at 34; and score 4 is -1.00×10 or -10 units from 50 or at 40 (see Table 24).

TABLE 24

COMPARISON OF *T*-SCORES AND STANDARD (*Z*) SCORES

(Data from Table 22)

Test Score	f	*T*-Scores	Standard (*Z*) Scores $M = 50$, $\sigma = 10$
10	1	74	75
9	4	67	69
8	6	61	63
7	10	56	57
6	8	52	52
5	13	48	46
4	18	41	40
3	2	28	34

$N = 62$

For test scores:

$M = 5.73$
$\sigma = 1.72$

Equation for converting test scores into standard scores (see p. 157)

$$\frac{X - 5.73}{1.72} = \frac{Z - 50}{10}$$

$$Z = \frac{10X}{1.72} - \frac{57.3}{1.72} + 50$$

$$Z = 5.82X - 33.3 + 50$$
$$Z = 5.82X + 16.7$$

The simplest plan for converting raw scores into σ-scores is to set up an equation as shown in Table 24. Here $\frac{X - 5.73}{1.72} = \frac{Z - 50}{10}$, that is, X-scores in the original distribution, expressed as σ-deviations from 5.73, are equal to Z-scores in the new distribution expressed as σ-deviations from 50. $Z = 5.82X + 16.7$, and on substituting our X's (i.e., 3, 4, 5, etc.) we obtain equivalent Z's (i.e., 34, 40, 46, etc.). These Z-scores correspond fairly closely to T-scores, and the more "normal" the original distribution the closer is the correspondence. The two kinds of scores are not interchangeable, however. With respect to the original scores, T-scores represent equivalent scores in a *normal distribution.* Standard or Z-scores, on the other hand, have the *same* form of distribution as the original scores, and are simply original scores expressed in σ-units. Z-scores represent the transformation we make when inches are changed into centimeters, or pounds are changed into kilograms. Both of these operations are "linear transformations,"* and involve no assumption as to form distribution.

(3) Percentile Scaling

In percentile scaling, a child who makes a certain score upon a test is given a percentile rank of 27, 36, or 77, say, in accordance with his position in the distribution. When the distribution of each of several tests has been drawn up, individual scores may be readily translated into percentile ranks. These ranks may then be compared directly, or combined to give a final percentile ranking. The method of computing percentiles has already been considered (p. 77). It is only necessary here, therefore, to show how percentile rankings may be compared, or combined into a final score.

Table 25 gives the percentile distributions for nine-year-olds

* When the equation connecting Z with X is that of a straight line (the general form of a straight line equation is $y = mx + b$), changing X's into Z's involves a "linear transformation."

upon three tests of the Pintner-Paterson series of performance tests.*

TABLE 25

PERCENTILE DISTRIBUTIONS FOR NINE-YEAR-OLDS ON THREE TESTS

Method of Combining the Percentile Ranks of a Single Individual

	Percentiles												
Tests	0	10	20	30	40	50	60	70	80	90	100	*S*'s Score	*S*'s Perc. Rank
Picture Completion	62	240	297	325	372	407	440	450	499	577	646	445	65
Substitution	219	190	173	158	152	141	133	126	121	109	80	126	70
Seguin Form-Board	34	24	21	20	18	18	17	16	15	15	13	17	60
Median Percentile Rank													65

The subject, a nine-year-old boy, made a score of 445 on the completion test which gives him a percentile rank of 65 (midway between 60 and 70). On the substitution test, a score of 126 gives him a percentile rank of 70; and on the Seguin formboard a score of 17 gives him a percentile rank of 60. The scores on tests two and three are in time units (seconds) so that the lowest score numerically represents the highest achievement.

The median of this subject's three percentile ranks is 65, which indicates that he stands somewhat above the median of children of his age in these tests. If this subject had been ten or eleven years old, percentile distributions for these ages would, of course, have been used. Percentile ranks may be combined directly when such derived scores are expressed in comparable units. Each test then has equal weight in the final score.

Percentile scales assume that the difference between ranks of 10 and 20 is the same as the difference between ranks of 40 and 50; that is, percentile differences are taken to be equal throughout the scale. This assumption holds strictly *only* when the distribution of scores is in the form of a *rectangle* rather than in the form of a *normal curve*. Figure 39 shows graphically the difference between the two types of distribution. The figure represents a rectangular distribution and a normal distribution of the same area plotted over it. The rectangular distribution

* Pintner, R., and Paterson, D. G., *A Scale of Performance Tests* (1925), pp. 189 and 197.

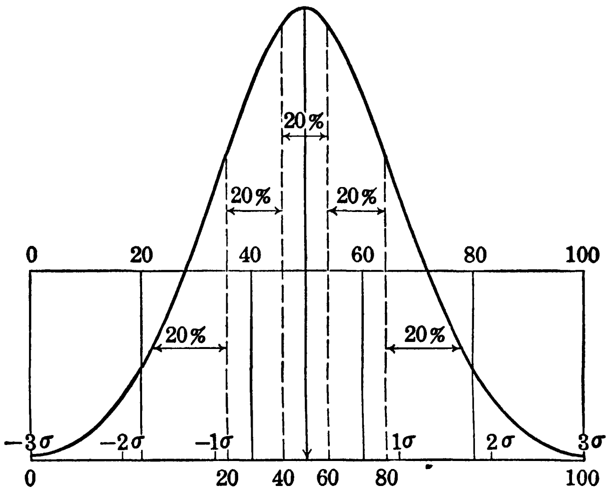


Fig. 39. To Illustrate the Position of the Same Five Percentiles in Rectangular and Normal Distributions.

has been divided into five equal parts or quintiles by taking successive fifths of the area. Along the top of the rectangle, a linear scale comprising five equal units is laid off. The width of each small rectangle is the same — the distances from 0 to 20, from 20 to 40, from 40 to 60, from 60 to 80, and from 80 to 100 are all equal. Now let us compare these equal percentile distances with the same percentile distances calculated from the normal curve. The first 20% of area, counted off from the extreme left of the normal curve, covers almost twice the distance along the baseline of the curve as is occupied by the first 20% of the rectangular distribution. This first 20% also covers about four times as much of the baseline as the third 20% (i.e., that from 40 to 60) in the normal curve. The baseline extent covered by the first 20% in the normal curve has been found in the following way: From Table 17 we find that the 30% of the area to the left of the mean extends from the mean to point $-.84\sigma$. Hence, the first 20% of the normal distribution falls between -3.00σ and $-.84\sigma$. The second 20% lies

between $-.84\sigma$ and $-.25\sigma$ since point $.25\sigma$ lies at a distance of 10% from the mean. The third 20% lies between $-.25\sigma$ and $.25\sigma$. The fourth and fifth 20%'s occupy the same relative positions in the upper half of the curve as the second and first 20%'s occupy in the lower half of the curve. It is clear that the steps from 0 to 20 and from 20 to 40 are not equal when measured along the baseline of the normal curve. Note that this inequality is relatively greater at the extremes of the distribution than it is around the mean.

Since most distributions of test scores tend to be normal or approximately so, equal percentile distances cannot usually be taken to represent equal steps in difficulty throughout the percentile scale. Between Q_1 and Q_3, percentile ranks are approximately equally spaced. Percentile ranks of a child in two different tests may be combined or averaged with little error when they fall between these limits. But percentile ranks *greater* than 75 or *less* than 25 should be combined, if at all, with full knowledge of their limitations.

III. The Transformation of Measures by Relative Position into Units of Amount

1. Product Scales. The Conversion of Judgments of Relative Merit into σ or PE Units

We have seen in the last section how test scores may be scaled on the principle that the σ-value determined from the percentage passing a given item is an acceptable index of difficulty. In scaling scores the assumption is made that ability is normally distributed from poor to good, and that performance may be scored quantitatively in terms of amount or time. It often happens, however, that the ability or trait in which we are interested is of such a nature that achievement cannot be expressed by a test score. This necessitates the construction of what are called "product scales." On such scales excellence of performance is evaluated by comparing an individual's production with various "standard productions" the values of which have been determined beforehand by a consensus of experts.

Handwriting, compositions, and drawing scales are well-known examples of product scales. The excellence of a person's penmanship, for example, can be determined by comparing a sample of his writing with various specimens of handwriting, the quality of which has been measured against some criterion.

Product scales are constructed on the principle that "equally often noticed differences" in quality are equal. If composition A, for example, is rated better than composition B by 75% of a group of competent judges, and composition X is rated better than composition Y by 75% of the same judges, then the difference between A and B is taken to be the same as the difference between X and Y (because equally often observed).

The assumption that "equally often noticed differences are equal" has been criticized* and is most doubtful when applied to the scaling of items at the extremes of the qualitative range. The variability of judgments upon extremely good or extremely poor specimens will ordinarily be less than the range of judgments made upon intermediate specimens. In most product scales the accurate measurement of these extreme specimens is, perhaps, not so important as is the accurate scaling of those items which constitute the main body of the scale. For this reason, the assumption that equally often noticed differences are equal will usually give scales which are as useful practically as those resulting from the use of more refined techniques.

Steps in constructing a product scale may be set down as follows:

(1) Collect a large number of samples of the product to be scaled (e.g., handwriting, drawings, jokes, pictures). These specimens should range by gradual stages from very poor to excellent.

(2) Persuade a number of competent persons to act as judges of the comparative excellence of the specimens. These judges are instructed to compare every specimen with every other

* Thurstone, L. L., "Equally Often Noticed Differences," *Journal of Educational Psychology*, 18 (1927), 289–293.

Thurstone, L. L., "Psychophysical Analysis," *American Journal of Psychology*, 38 (1927), 368–389.

specimen, so that a consensus may be obtained on each. The order of merit method, the paired comparisons method, or some variation of these, should ordinarily be employed here, as these experimental techniques provide a systematic attack upon the problem of ranking samples for excellence.*

(3) Reduce the number of times each specimen is ranked above each other specimen to percentage terms, and express these percents as σ-distances between each pair of specimens. To illustrate, if drawing A is judged better than drawing B by 65% of the group, $A - B = .39\sigma$; if B is judged better than C by 77%, $B - C = .74\sigma$. These σ-differences are read from Table 17 and are found in the following way: If a sample is judged better than another by just 50%, there is no observable difference between the two and their σ-difference is zero. But if A is judged better than B by 65%, the difference between A and B (in excess of chance) is 15%, which from Table 17 corresponds to a σ-difference of .39. In exactly the same way the difference between B and C (in excess of chance) is 27%, which corresponds to a σ-difference of .74. Figure 40 shows graphically how percentage differences can be converted into σ-differences. The distributions of judgments upon A, B, and C are assumed to be normal and are taken to be equal in range and variability. The mean value of A (its scale value) is $.39\sigma$ above the mean value of B, whose mean value is, in turn, $.74\sigma$ above the mean value of C.

(4) Determine a difference for each pair of specimens, and express each item finally selected for the scale as so many σ-units from the arbitrary zero. The procedure may be illustrated by two items, numbers eight and nine, taken from the Hillegas Composition Scale.† Hillegas had each of 202 judges arrange a number of English compositions in

* Woodworth, R. S., *Experimental Psychology* (1938), pp. 372–378.

† Hillegas, Milo B., *A Scale for the Measurement of Quality in English Composition by Young People*, Teachers College Record, 13 (1912), 4, 5–55.

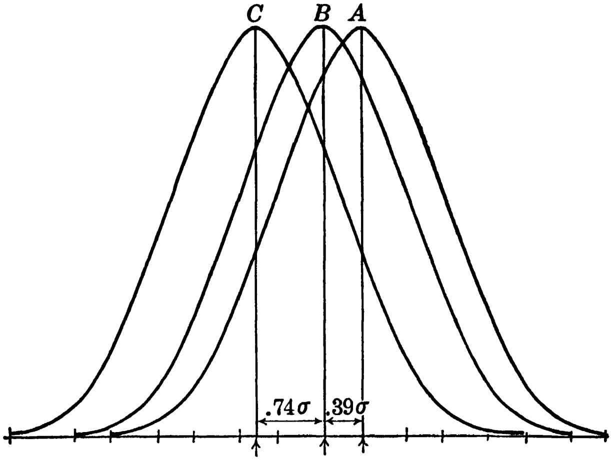


FIG. 40. To Illustrate σ-Scale Differences between Specimens A, B, and C. The distributions of judgments on the three specimens are taken to be normal, and equal in range and variability.

order of merit. An artificial composition was selected as being of just zero merit, and assigned the value of 0 on the scale. Of the 202 judges, 136 or 67.33% ranked specimen nine as better than specimen eight. From Table 18, we find that a percentage difference of 67.33 indicates a *PE* difference of .65, and this value expresses the amount by which nine is better than eight. The value of specimen eight had already been found to be 7.72*PE* above the zero point on the scale. Hence, specimen nine is 7.72 + .65 or 8.37*PE* above the zero composition. The values of the nine compositions on the Hillegas Scale as measured in *PE* units from the zero composition are 1.83, 2.60, 3.69, 4.74, 5.85, 6.75, 7.72, 8.37, and 9.37. Note that the steps on the scale are fairly regular and are about 1*PE* apart.

2. The Transformation of Qualitative Data into Numerical Scores

It is possible to represent many kinds of qualitative data in quantitative terms, if we can assume that measures of the trait

or ability sampled by our data are normally distributed. Two examples, which are typical of many, will be given by way of illustration.

(1) The Scaling of Answers to a Questionnaire

The answers to the queries or statements in most questionnaires admit of several possible replies, such as Yes, No, ?; or Most, Many, Some, Few, No; or there are four or five answers one of which is to be checked. It is often desirable to "weight" these different alternatives in accordance with the degree of divergence from the "typical answer" which they indicate. Let us first assume that the attitude or other personality trait expressed in answering a given proposition is normally distributed. From the percentage who accept each alternative answer to a question or statement, we may then find a σ-equivalent, which will express the value or weight to be given that answer. Likert's* Internationalism Scale furnishes an example of this scaling technique. This questionnaire contains twenty-four statements upon each of which the subject is requested to give an opinion. Approval or disapproval of any statement is indicated by checking one of five possibilities "strongly approve," "approve," "undecided," "disapprove," and "strongly disapprove." The method of scaling as applied to statement No. 16 on the Internationalism Scale is shown in Table 26 on page 165. This statement reads as follows:

16. All men who have the opportunity should enlist in the Citizens' Military Training Camps.
    Strongly approve Approve Undecided Disapprove Strongly disapprove

The percentage selecting each of the possible answers is shown in the table. Below the percent entries are the σ-equivalents assigned to each alternative on the assumption that opinion on the question is normally distributed — that few will whole-

* Likert, R., *A Technique for the Measurement of Attitudes*, Archives of Psychology, No. 140 (1932).

TABLE 26

DATA FOR STATEMENT NO. 16 OF THE INTERNATIONALISM SCALE

Answers	Strongly Approve	Approve	Undecided	Disapprove	Strongly Disapprove
Percent checking	13	43	21	13	10
Equivalent σ-values	-1.63	$-.43$	.43	.99	1.76
Z-scores	34	46	54	60	68

heartedly agree or disagree, and many take intermediate views. The σ-values in Table 26 have been obtained from Table 27 (p. 167) in the following way: Reading down the first column

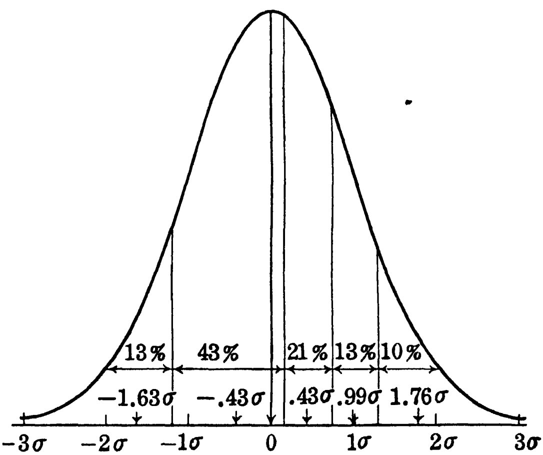


FIG. 41. To Illustrate the Scaling of the Five Possible Answers to Statement 16 on Likert's Internationalism Scale.

headed 0, we find that beginning at the upper extreme of the normal distribution, the highest 10% has an average σ-distance from the mean of 1.76. Said differently, the mean of the 10% of cases at the upper extreme of the normal curve is at a distance of 1.76σ from the mean of the whole distribution. Hence, the answer "strongly disapprove" is given a σ-equivalent of 1.76 (see Fig. 41).

To find the σ-value for the answer "disapprove," we select

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	270	218	196	181	170	160	151	144	137	131	125	120	115	110	106	102	97	94	90	86	82	79	76	72
2	244	207	189	175	165	156	148	141	134	128	122	118	112	108	104	99	95	92	88	84	81	77	74	71
3	228	198	182	170	160	152	144	137	131	125	120	115	110	106	102	97	94	90	86	82	79	76	72	69
4	216	191	177	165	156	148	141	134	128	123	118	113	108	104	100	96	92	88	84	81	77	74	71	67
5	210	185	172	161	152	145	138	131	126	120	115	111	106	102	98	94	90	86	82	79	76	72	69	66
6	199	179	167	157	149	141	135	129	123	118	113	108	104	100	96	92	88	84	81	77	74	71	68	64
7	192	174	163	153	145	138	132	126	121	116	111	106	102	98	94	90	86	83	79	76	72	69	66	63
8	186	170	159	150	142	135	128	124	118	113	109	104	100	96	92	88	84	81	77	74	71	68	64	61
9	181	165	155	147	139	133	126	121	116	111	106	102	98	94	90	86	83	79	76	73	69	66	63	60
10	176	161	151	143	136	130	124	119	114	109	104	100	96	92	88	85	81	78	74	71	68	65	62	59
11	171	158	148	140	134	127	122	116	111	107	102	98	94	90	87	83	79	76	73	69	66	63	60	57
12	167	154	145	138	131	125	119	114	109	105	100	96	92	89	85	81	78	74	71	68	65	62	59	56
13	163	151	142	135	128	122	117	112	107	103	99	94	91	87	83	80	76	73	70	66	63	60	57	54
14	159	147	139	132	126	120	115	110	105	101	97	93	89	85	81	78	75	71	68	65	62	59	56	53
15	156	144	136	129	123	118	113	108	103	99	95	91	87	83	80	76	73	70	66	63	60	57	54	51
16	152	141	134	127	121	116	111	106	101	97	93	89	85	82	78	75	71	68	65	62	59	56	53	50
17	149	139	131	125	119	113	109	104	99	95	91	87	84	80	77	73	70	67	64	60	57	54	52	49
18	146	136	129	122	117	111	106	102	98	93	89	86	82	78	75	72	68	65	62	59	56	53	50	47
19	143	133	126	120	114	109	105	100	96	92	88	84	80	77	73	70	67	64	61	58	55	52	49	46
20	140	131	124	118	112	107	103	98	94	90	86	82	79	75	72	69	65	62	59	56	53	50	47	45
21	137	128	121	116	110	105	101	96	92	88	84	81	77	74	70	67	64	60	58	55	52	49	46	43
22	135	126	119	113	108	103	99	95	90	87	83	79	76	72	69	66	62	59	56	53	50	48	45	42
23	132	124	117	111	106	101	97	92	89	85	81	78	74	71	67	64	61	58	55	52	49	46	43	41
24	130	121	115	109	104	100	95	91	87	83	80	76	73	69	66	63	60	57	54	51	48	45	42	39
25	127	119	113	107	102	98	93	89	85	82	78	74	71	68	64	61	58	55	52	49	46	43	41	38
26	125	117	111	105	101	96	92	88	84	80	76	73	70	66	63	60	57	54	51	48	45	42	39	37
27	123	115	109	104	99	94	90	86	82	78	75	71	68	65	62	58	55	52	49	46	44	41	38	35
28	120	113	107	102	97	92	88	84	80	77	73	70	67	63	60	57	54	51	48	45	42	39	37	
29	118	111	105	100	95	91	87	83	79	75	72	68	65	62	59	56	53	50	47	44	41	38		
30	116	109	103	98	93	89	85	81	77	74	70	67	64	60	57	54	51	48	45	42	40			
31	114	107	101	96	92	87	83	79	76	72	69	65	62	59	56	53	50	47	44	41				
32	112	105	99	94	90	86	82	78	74	71	67	64	61	58	54	51	48	46	43					
33	110	103	98	93	88	84	80	76	73	69	66	63	59	56	53	50	47	44						
34	108	101	96	91	86	82	79	75	71	68	64	61	58	55	52	49	46							
35	106	99	94	89	85	81	77	73	70	66	63	60	56	53	50	47								
36	104	97	92	88	83	80	75	72	68	65	61	58	55	52	49									
37	102	96	91	86	82	78	74	70	67	63	60	57	54	51										
38	100	94	89	84	80	76	72	69	65	62	59	55	52											
39	98	92	87	83	79	75	71	67	64	61	57	54												
40	97	91	86	81	77	73	69	66	62	59	56													
41	95	89	84	80	75	72	68	64	61	58														
42	93	87	82	78	74	70	66	63	60															
43	91	85	81	76	72	69	65	62																
44	90	84	79	75	71	67	64																	
45	88	82	78	73	69	66																		
46	86	81	76	72	68																			
47	85	79	75	70																				
48	83	78	73																					
49	81	76																						
50	80																							

	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
1	69	66	63	60	57	54	51	48	45	43	40	37	35	32	29	27	24	21	19	16	14	11	09	06	04	01
2	67	64	61	58	55	52	50	47	44	41	39	36	33	31	28	25	23	20	18	15	13	10	08	05	03	
3	66	63	60	57	54	51	48	45	43	40	37	35	32	29	27	24	21	19	16	14	11	09	06	05		
4	64	61	58	55	52	50	47	44	41	39	36	33	31	28	25	23	20	18	15	13	10	08	05			
5	63	60	57	54	51	48	45	43	40	37	35	32	29	27	24	21	19	16	14	11	09	06				
6	61	58	55	53	50	47	44	41	39	36	33	31	28	25	23	20	18	15	13	10	08					
7	60	57	54	51	48	45	43	40	37	35	32	29	27	24	21	19	16	14	11	09						
8	58	55	52	50	47	44	41	39	36	33	31	28	25	23	20	18	15	13	10							
9	57	54	51	48	46	43	40	37	35	32	29	27	24	21	19	16	14	11								
10	56	53	50	47	44	41	39	36	33	31	28	25	23	20	18	15	13									
11	54	51	48	46	43	40	37	35	32	29	27	24	22	19	16	14										
12	53	50	47	44	41	39	36	33	31	28	25	23	20	18	15											
13	51	48	46	43	40	37	35	32	29	27	24	22	19	16												
14	50	47	44	42	39	36	33	31	28	25	23	20	18													
15	49	46	43	40	37	35	32	29	27	24	22	19														
16	47	44	42	39	36	33	31	28	26	23	20															
17	46	43	40	37	35	32	29	27	24	22																
18	44	42	39	36	33	31	28	26	23																	
19	43	40	38	35	32	30	27	24																		
20	42	39	36	34	31	28	26																			
21	40	38	35	32	30	27																				
22	39	36	34	31	28																					
23	38	35	32	30																						
24	36	34	31																							
25	35	32																								
26	34																									

TABLE 27

Average distance from the mean, in terms of σ, of each single percentage of a normal distribution. Figures along the top of the table represent percentages of area from either extreme. Figures down the side of the table represent percentages measured from given points in the distribution.

Examples: The average distance from the mean of the highest 10% of a normally distributed group is 1.76σ (entry opposite 10 in first column). The average distance from the mean of the *next* 20% is $.86\sigma$ (entry opposite 20 in column headed 10). The average distance from the mean of the *next* 30% is

$$\frac{.26 \times .20 + (-\ .13 \times .10)}{.30}$$

or $.13\sigma$ (20% lie to right of mean and 10% to left, see p. 165).

the column headed .10 and running down the column take the entry opposite 13, namely, .99. This means that when 10% of the distribution reading from the upper extreme have been accounted for, the average distance from the mean of the next 13% is .99σ. Reference to Figure 41 will make this clearer. Now from the column headed 23 (13% + 10% "used up" or accounted for), we find entry .43 opposite 21. This means that when the 23% at the upper end of the distribution have been cut off, the mean σ-distance from the general mean of the next 21% is .43σ, which becomes the weight of the preference "undecided." The weight of the fourth answer "approve" must be found by a slightly different process. Since a total of 44% from the upper end of the distribution have now been accounted for, 6% of the 43% who marked "approve" will lie to the *right* of the mean, and 37% to the *left* of the mean, as shown in Figure 41. From the column headed 44 in Table 27, we take .08 (entry opposite 6%) which is the average distance from the general mean of the 6% lying just above the mean. Then from the column headed 13 (50% − 37%) we take entry .51 (now − .51) opposite 37%, as the mean distance from the general mean of the 37% just below the mean. The algebraic sum $\frac{-.51 \times .37 + .08 \times .06}{.43} = -.43$, which is the weight assigned to the preference "approve." The 13% left, those marking "strongly approve," occupy the 13% at the extreme (low end) of the curve. Returning to the column headed 0, we find that the mean distance from the general mean of the 13% at the extreme of the distribution is − 1.63σ.

In order to avoid negative values, each σ-weight in Table 26 can be expressed as a σ-distance from − 3.00σ (or − 5.00σ). If referred to − 3.00σ, the weights become in order 1.37, 2.57, 3.43, 3.99, and 4.76. Dropping decimals, and taking the first two digits, we could also assign weights of 14, 26, 34, 40, and 48. Again each σ-value in Table 26 may be expressed as a Z-score. In a distribution the mean of which is 50 and the σ 10, the category "strongly approve" is − 16(− 1.63 × 10)

from the mean of 50, or at 34. Category "approve" is $-4(-.43 \times 10)$ from 50 or at 46. The other three categories have Z-scores of 54, 60, and 68.

When all of the twenty-four statements on the Internationalism Scale have been scaled as shown above, a person's "score" (his attitude toward internationalism in general) is found by adding up the weights assigned to the various preferences which he has selected. An individual whose opinions are extreme, e.g., who tends strongly to disapprove many statements, will receive a proportionally larger total score when the choices are σ-scaled, than he would receive if the five possibilities were assigned arbitrary weights of 1, 2, 3, 4, and 5. Likert has shown, however, that σ-scaling yields results which, *for the test as a whole*, are little if any more reliable or more discriminatory than the results obtained when the five answers are scored simply 1, 2, 3, 4, and 5. This virtual equality of scaling and rule-of-thumb method is a rather familiar finding in mental measurement. In the present instance, it probably arises from the fact that the greater differentiation which the σ-scaling technique provides for *single* items is lost in the process of adding or averaging the score weights from many items. A real advantage of σ-scaling is that the units of the scale are equal and may be compared from item to item or from scale to scale. Also, σ-scaling gives a more accurate picture of the extent to which extreme or biased opinions on a given question are divergent from the typical opinion than does the arbitrary weighting method.

(2) The Scaling of Judgments or Ratings

In many psychological problems, individuals are rated or ranked for their possession of characteristics or attributes not readily measured by tests. Honesty, interest in one's work, tactfulness, originality, are illustrations of such traits. Suppose that two teachers A and B have rated a group of forty pupils for "social responsibility" on a 5-point scale. A rating of 1 means that the trait is possessed in marked degree, a rating of 5 that it is almost if not entirely absent, and ratings of 2, 3,

and 4 indicate intermediate degrees. Assume that the percentage of children assigned each rating is as follows:

Social Responsibility

Rating	A	B
1	10%	20%
2	15%	40%
3	50%	20%
4	20%	10%
5	5%	10%

It is obvious that B rates more leniently than A, so that a rating of 1 by B does not represent the same degree of "social responsibility" as a rating of 1 by A. Can we assign "weights" or numerical scores so as to make the ratings of the two teachers comparable? The answer is "yes," provided we can assume that the distribution of the trait "social responsibility" is normal, and that one teacher is as competent a judge as the other. From Table 27, we may read σ-equivalents to the percents given each rating by A and B as follows:

Rating	A	B
1	1.76	1.40
2	.95	.27
3	.00	− .53
4	− 1.07	− 1.04
5	− 2.10	− 1.76

These σ-values are read from Table 27 in exactly the same way as were the σ-equivalents in the previous problem (p. 165). If we assume -3.00σ as an arbitrary reference point, the σ-values for the ratings of A and B all become positive:

Rating	A	B
1	4.76	4.40
2	3.95	3.27
3	3.00	2.47
4	1.93	1.96
5	.90	1.24

Dropping decimals, and taking only the first two digits, A's and B's ratings become:

Rating	A	B
1	48	44
2	40	33
3	30	25
4	19	20
5	9	12

Or, expressed as Z-scores in a distribution with a mean of 50 and a σ of 10,

Rating	A	B
1	68	64
2	60	53
3	50	45
4	39	40
5	29	32

It is possible to combine the ratings of A and B by adding or by averaging them. If a child receives a rating of "4" by A and a rating of "2" by B, his combined or average rating would be $\frac{-1.07 + .27}{2}$ or $-.40$; $\frac{1.93 + 3.27}{2}$ or 2.60; $\frac{19 + 33}{2}$ or 26; $\frac{39 + 53}{2}$ or 46.

Table 27 will prove valuable in enabling one to transmute many kinds of qualitative data into quantitative terms or scores. Almost any attribute upon which relative judgments can be obtained may be assigned scores in a normal distribution in terms of the σ of the judgments.

(3) Changing Order of Merit Ranks into Numerical Scores

It is often desirable to transmute orders of merit into units of amount or "scores." This may be done by means of tables, if we are justified in assuming normality for the trait in which the ranking has been made. To illustrate, suppose that fifteen salesmen have been ranked in order of merit for selling

efficiency, the most efficient salesman being ranked 1, the least efficient being ranked 15. If we are justified in assuming that "selling efficiency" follows the normal probability curve, we can, with the aid of Table 28 (p. 173), assign to each man a "selling score" on a scale of 100 points. Such a score will probably represent his ability as a salesman better than will a rank of 2, 6, or 14. The problem may be stated specifically as follows:

Example (*1*) Given fifteen salesmen, ranked in order of merit by their sales manager, to transmute these rankings into scores on a scale of 100 points.

First, by means of the formula

$$\textit{Percent position} = \frac{100(R - .5)}{N} \tag{21}$$

(*formula for transmuting ranks into percents*)

in which R is the rank of the individual in the series* and N is the number of individuals ranked, determine the "percent position" of each man. Then from these percent positions read the man's score on a scale of 100 points from Table 28. Salesman A, who ranks No. 1, has a percent position of $\frac{100(1 - .5)}{15}$ or 3.33, and his score from Table 28 is 85 (finer interpolation unnecessary). Salesman B, who ranks No. 2, has a percent position of $\frac{100(2 - .5)}{15}$ or 10, and his score, accordingly, is 75. The scores of the other salesmen, found in exactly the same way, are given in the table on page 174.

It has been frequently pointed out that the assumption of normality in a trait implies that differences at the extremes of the trait are relatively much greater than differences around the mean. This is clearly brought out in the next table; for, while all differences in the order of merit series equal 1, the differences

* A rank is an interval on a scale; .5 is subtracted from each R because its midpoint best represents an interval. E.g., $R = 5$ is the 5th interval, namely 4–5, and 4.5 (or $5 - .5$) is the midpoint.

TABLE 28

THE TRANSMUTATION OF ORDERS OF MERIT INTO UNITS OF AMOUNT OR "SCORES" *

Example: If $N = 25$, and $R = 3$, Percent Position is $\frac{100(3 - .5)}{25}$ or 10 (formula 21) and from the table, the equivalent rank is 75, on a scale of 100 points.

Percent	Score	Percent	Score	Percent	Score
.09	99	22.32	65	83.31	31
.20	98	23.88	64	84.56	30
.32	97	25.48	63	85.75	29
.45	96	27.15	62	86.89	28
.61	95	28.86	61	87.96	27
.78	94	30.61	60	88.97	26
.97	93	32.42	59	89.94	25
1.18	92	34.25	58	90.83	24
1.42	91	36.15	57	91.67	23
1.68	90	38.06	56	92.45	22
1.96	89	40.01	55	93.19	21
2.28	88	41.97	54	93.86	20
2.63	87	43.97	53	94.49	19
3.01	86	45.97	52	95.08	18
3.43	85	47.98	51	95.62	17
3.89	84	50.00	50	96.11	16
4.38	83	52.02	49	96.57	15
4.92	82	54.03	48	96.99	14
5.51	81	56.03	47	97.37	13
6.14	80	58.03	46	97.72	12
6.81	79	59.99	45	98.04	11
7.55	78	61.94	44	98.32	10
8.33	77	63.85	43	98.58	9
9.17	76	65.75	42	98.82	8
10.06	75	67.48	41	99.03	7
11.03	74	69.39	40	99.22	6
12.04	73	71.14	39	99.39	5
13.11	72	72.85	38	99.55	4
14.25	71	74.52	37	99.68	3
15.44	70	76.12	36	99.80	2
16.69	69	77.68	35	99.91	1
18.01	68	79.17	34	100.00	0
19.39	67	80.61	33		
20.93	66	81.99	32		

between the transmuted scores vary considerably. The greatest differences are found at the ends of the series, the smallest in the middle. For example, the difference in score between A and B or between N and O is three times the difference between G

* From Hull, C. L., *The Computation of Pearson's* r *from Ranked Data,* Journal of Applied Psychology (1922), 6, pp. 385–390.

and H. Clearly it is three times as hard for a salesman to improve sufficiently to move from second to first place, as it is for him to improve sufficiently to move from eighth to seventh place.

The percentile ranks (*PR*'s) of our fifteen salesmen are also given in the table for comparison with the normal curve "scores." *PR*'s were calculated by the method given on page 80. Note that the steps between *PR*'s are all equal;

Salesmen	Order of Merit Ranks	Percent Position (Table 28)	Score (Scale 100)	*PR*
A	1	3.33	85	97
B	2	10.00	75	90
C	3	16.67	69	83
D	4	23.33	64	77
E	5	30.00	60	70
F	6	36.67	57	63
G	7	43.33	53	57
H	8	50.00	50	50
I	9	56.67	47	43
J	10	63.33	43	37
K	11	70.00	40	30
L	12	76.67	36	23
M	13	83.33	31	17
N	14	90.00	25	10
O	15	96.67	15	3

there are no differences between the *PR*'s at intermediate and at extreme positions. Both ranks and *PR*'s assume that the distribution of ability is rectangular rather than normal in form (p. 159). Equal slices of area correspond directly to equal distances along the baseline.

Another use to which Table 28 may be put is in the combination of incomplete order of merit rankings. To illustrate:

Example (2) Six persons, A, B, C, D, E, and F, are to be ranked for honesty by three judges. Judge 1 knows all six well enough to rank them; Judge 2 knows only three well enough to rank them; and Judge 3 knows four well enough to rank them. Can we obtain a fair composite order of merit ranking for all six persons by combining these three sets of rankings, two of which are incomplete?

We may tabulate our data as follows:

	Persons					
	A	B	C	D	E	F
Judge 1's ranking	1	2	3	4	5	6
Judge 2's ranking		2		1		3
Judge 3's ranking	2		1		3	4

It seems fair that A should get more credit for ranking first in a list of six, than D for ranking first in a list of three, or C for ranking first in a list of four. In the order of merit ratings, all three individuals are given the same rank. But when we assign scores to each person, in accordance with his position in the list, by means of formula 21 and Table 28, A gets 77 for his first place, D gets 69 for his, and C gets 73 for his. See table below:

	Persons					
	A	B	C	D	E	F
Judge 1's ranking	1	2	3	4	5	6
score	77	63	54	46	37	23
Judge 2's ranking		2		1		3
score		50		69		31
Judge 3's ranking	2		1		3	4
score	56		73		44	27
Sum of scores	133	113	127	115	81	81
Mean	67	57	64	58	41	27
Order of Merit	1	4	2	3	5	6

All of the ratings have been transmuted as shown in example (1) above. Separate scores may be combined and averaged to give the final order of merit shown in the table.

By means of formula 21 and Table 28 it is possible to transmute any set of ranks into scores, if we may assume a normal distribution in the trait for which the ranking is made. The method is useful in the case of those attributes which are not easily measured by ordinary methods, but for which individuals may be arranged in order of merit, as, for example, athletic ability, personality, beauty, and the like. It is also valuable in correlation problems when the only available criterion* of a given ability or aptitude is a set of ranks. Trans-

* For definition of a criterion, see Chapter XII, p. 394.

muted scores may be combined or averaged like other test scores.

A word of explanation may be added with regard to Table 28. This table represents a normal frequency distribution which has been cut off at $\pm 2.5\sigma$. The baseline of the curve is 5σ, therefore, and may conveniently be divided into 100 parts, each $.05\sigma$ long. The first $.05\sigma$ from the upper limit of the curve takes in .09 of 1% of the distribution and is scored 99 on a scale of 100. The next $.05\sigma$ ($.10\sigma$ from the upper end of the curve) takes in .20 of 1% of the entire distribution and is scored 98. In each case, the percent position gives the fractional part of the normal distribution which lies to the right of (above) the given

PROBLEMS

1. In a sample of 1000 cases the mean of a certain test is 14.40 and σ is 2.50. Assuming normality of distribution
   (a) How many individuals score between 12 and 16?
   (b) How many score above 18? below 8?
   (c) What are the chances that any individual selected at random will score above 15?
2. In a distribution of 100 cases, the median is 29.74 and the Q is 3.18. Assuming normality
   (a) What percent of the cases lie between 24 and 25?
   (b) What limits include the middle 60%?
   (c) What limits include the lowest 5%?
3. In a certain achievement test, the seventh-grade median is 28.00, with a Q of 4.80; and the eighth-grade median is 31.60 with a Q of 4.00. What percent of the seventh grade is above the median of the eighth grade? What percent of the eighth grade is below the median of the seventh grade?
4. Two years ago a group of twelve-year-olds had a reading ability expressed by a mean score of 40.00 and a σ of 3.60; and a composition ability expressed by a mean of 62.00 and a σ of 9.60. Today the group has gained 12 points in reading and 10.8 points in composition. How many times greater is the gain in reading than the gain in composition?

5. In Problem 1, Chapter IV, we computed directly from the distribution the percent of Group A which exceeds the median of Group B. Compare this value with the percentage of overlapping obtained on the assumption of normality in Group A.

6. Four problems A, B, C, and D, have been solved by 50%, 60%, 70%, and 80%, respectively, of a large group. Compare the difference in difficulty between A and B with the difference in difficulty between C and D.

7. In a certain college, ten grades, A+, A, A−; B+, B, B−; C+, C, C−; and D, are assigned. If ability in mathematics is distributed normally, how many students in a group of 500 freshmen should receive each grade?

8. Five problems are passed by 15%, 34%, 50%, 62%, and 80%, respectively, of a large unselected group. If the zero point of ability in this test is taken to be at -3σ, what is the σ-value of each problem as measured from this point?

9. (*a*) Locate the deciles in a normal distribution in the following way. Beginning at -3σ, count off successive 10%'s of area up to $+3\sigma$. Tabulate the σ-values of the points which mark off the limits of each division. For example, the limits of the first 10% from -3σ are -3.00σ and -1.28σ (see Table 17, p. 115.) Label these points in order from -3σ as .10, .20, etc. Now compare the distances in terms of σ between successive ten percent points. Explain why these distances are unequal.
   (*b*) Divide the baseline of the normal probability curve (take as 6σ) into ten equal parts, and erect a perpendicular at each point of division. Compute the percentage of total area comprised by each division. Are these percents of area equal? If not, explain why. Compare these percents with those found in (*a*).

10. In a large group of competent judges, 88% rank composition A as better than composition B; 65% rank B as better than C. If C is known to have a *PE* value of 3.50 as measured from the "zero composition," i.e., the composition of just zero merit, what are the *PE* values of B and A as measured from this zero point?

11. Twenty-five men on a football squad are ranked by the coach in order of merit from 1 to 25 for all-around playing ability. On the assumption that general playing ability is normally distributed,

transmute these ranks into "scores" on a scale of 100 points. Compare these scores with the PR's of the ranks.

12. On an Occupational Interest Blank, each occupation is followed by five symbols, L! L ? D D!, which denote different degrees of "liking" and "disliking." The answers to one item are distributed as follows:

L!	L	?	D	D!
8%	20%	38%	24%	10%

(*a*) By means of Table 27 convert these percents into σ-units.
(*b*) Express each σ-value as a distance from "zero," taken at -3σ, and multiply by 10 throughout.
(*c*) Express each σ-value as a Z-score in a distribution of mean 50, σ 10.

13. Letter grades are assigned three classes by their teachers in English, history, and mathematics, as follows:

Mark	English	History	Mathematics
A	25	11	6
B	21	24	15
C	32	20	25
D	6	8	20
F	1	2	8
	85	65	74

(*a*) Express each distribution of grades in percents, and by means of Table 27 transform these percents into σ-values.
(*b*) Change these σ-values into 2-digit numbers and into Z-scores following the method on page 171.
(*c*) Find average grades [from (b)] for the following students:

Student	English	History	Mathematics
S. H.	A	B	C
F. M.	C	B	A
D. B.	B	D	F

14. Calculate *T*-scores in the following problem:

Scores	*f*	Percent below given score Plus One-half Reaching	T-score
91	2	.995	76
90	4	.980	71
89	6		
88	20		
87	24		
86	28		
85	40		
84	36		
83	24		
82	12		
81	4		
	200		

(The first two *T*-scores have been entered.)

Answers

1. (*a*) 570
   (*b*) 75; 5
   (*c*) 41 in 100
2. (*a*) 5%
   (*b*) 33.72 and 25.76
   (*c*) 21.95 and lowest score in the distribution
3. 31%; 27%
4. Three times as great.
5. 39% as compared with 42%.
6. Difference between A and B is $.25\sigma$; between C and D, $.32\sigma$.
7. Grades:

Grades:	A+	A	A−	B+	B	B−	C+	C	C−	D
Students Receiving:	3	14	40	80	113	113	80	40	14	3

8. In order: 4.04; 3.41; 3.00; 2.69; 2.16.
9. (*a*)

	.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
	− 3.00	− 1.28	− .84	− .52	− .25	0	.25	.52	.84	1.28	3.00
Diffs:		1.72	.44	.32	.27	.25	.25	.27	.32	.44	1.72

   (*b*) Percents of area in order: .68; 2.77; 7.92; 15.92; 22.57; 22.57; 15.92; 7.92; 2.77; .68.

10. B, 4.05*PE*; A, 5.80*PE*.

11.

Rank:	1	2	3	4	5	6	7	8	9	10	11	12	13
Score:	89	80	75	71	68	65	63	60	58	56	54	52	50
PR's:	98	94	90	86	82	78	74	70	66	62	58	54	50
Rank:	14	15	16	17	18	19	20	21	22	23	24	25	
Score:	48	46	44	42	40	37	35	32	29	25	20	11	
PR's:	46	42	38	34	30	26	22	18	14	10	6	2	

12.

	L!	L	?	D	D!
(*a*)	− 1.86	− .94	− .08	.80	1.76
(*b*)	11	21	29	38	48
(*c*)	31	41	49	58	68

13.

		F	D	C	B	A
(*a*)	English	− 2.70	− 1.74	− .65	.22	1.18
	History	− 2.28	− 1.38	− .53	.39	1.49
	Math.	− 1.71	− .71	.13	.94	1.86

(*b*)

	English		History		Mathematics	
	-3.00σ	Z	-3.00σ	Z	-3.00σ	Z
A	42	62	45	65	49	69
B	32	52	34	54	39	59
C	24	44	25	45	31	51
D	13	33	16	36	23	43
F	3	23	7	27	13	33

(*c*) S. H., 36 or 56; F. M., 36 or 56; D. B., 20 or 40.

14. *T*-scores:
76, 71, 67, 62, 58, 54, 49, 44, 39, 34, 27

CHAPTER VII

SAMPLING AND RELIABILITY

I. The Meaning of Reliability

The "true" mean or the "true" σ of any set of measurements (of height, mechanical aptitude, or intelligence, for example) is that value found by taking into account the scores made by *all* of the members of some defined group (called the *population*). It is rarely if ever possible to measure all of the individuals in a given population, say all of the ten-year-old boys in New York city. Hence we must usually be content to deal with "samples" drawn from our population, and owing to slight differences in the composition of these samples, means and σ's may be somewhat larger or somewhat smaller than their corresponding population values. True and obtained measures are referred to, respectively, as population *parameters* and sample *statistics*.* Sample statistics are always *estimates* of their population counterparts; and the accuracy of this estimate is a measure of the *reliability* of the statistic.

Although we may not be able to determine the parameters (true values) themselves we can compute limits within which the true mean or some other statistic may, with a certain degree of confidence, be expected to lie. As we shall see later this range, which may be large or small, serves as a useful index of the reliability or dependability of the calculated statistic. Whenever we have calculated a statistic then, we must ask ourselves these questions: "How reliable is my answer?" "How well does this mean or σ represent the true value which I should

* A statistic is any measure calculated from a sample as, for example, the mean or *SD*.

get by taking into account the entire population from which my sample was drawn?" The purpose of this chapter is to present methods which will enable us to answer these questions. The reliability of measures of central tendency will be first considered; then the reliability of measures of variability and of certain other important statistics; and finally the reliability of the differences between obtained measures.

II. The Reliability of Measures of Central Tendency

1. The Reliability of the Mean

(1) The Standard Error (SE) of the Mean (σ_M)

What is meant by the reliability of the mean can best be seen by examining the factors upon which the stability of this measure depends. Suppose that we wish to know the mean ability of college freshmen in the United States as shown by their scores upon the American Council Psychological Examination. To measure the achievement of college freshmen *in general* would require in strict logic that we test *all* of the freshmen in the United States. But this is obviously a stupendous task, and we must perforce be satisfied with taking the records of as *large* and as *representative** a sample of freshmen as we can find. This means that we cannot use freshmen from only a single institution or from only one section of the country; and that we must guard against selecting only those with high, or only those with low, scholastic records. The more successful we are in getting an "unselected" group, the more representative this group will be of all freshmen in the country. Evidently, therefore, the reliability of a mean depends for one thing upon how impartially we have chosen our sample.

Given an adequate sample, the reliability of a mean can be shown to depend mathematically† upon *two* characteristics of the distribution: (1) the number of cases (N) and (2) the variability or spread of the measures.

* For further discussion of sampling, see pp. 222–227.
† Kelley, T. L., *Statistical Method* (1923), pp. 82–83.

(*a*) It is clear that the number of cases must influence the stability of a mean, since the addition of even one extra measure to a series will change the mean unless the additional case happens to coincide with the mean exactly. Moreover, the addition of one score to a set of ten scores will effect a greater change in the obtained mean than the addition of one score to a set of 1000 scores, as each case counts for less in the larger group. It can be shown mathematically, as well as experimentally,* that the reliability of a sample mean will increase, not in proportion to the number of measures upon which it is based, but in proportion to the *square root* of the number of measures. The mean obtained from twenty-five scores, for example, is not twenty-five times, but $\sqrt{25}$ or five times, as reliable as a single measure. And a mean based upon thirty-six cases is not four times as reliable as a mean based upon nine cases, but only twice as reliable — since $\sqrt{36}$ divided by $\sqrt{9}$ equals 2.

(*b*) Reliability of a mean also depends upon the variability of the separate measures around the mean. If the σ of the distribution is large, the separate measures tend to scatter widely, and we are unable to say where those cases in the population which we have not measured will most probably fall — whether they will be close to, or far from, the mean. On the other hand, if the σ is small, we may be fairly certain that unmeasured cases will fall close to the mean. The reliability of an obtained mean, therefore, varies with the size of the σ; as σ increases, the reliability decreases.

To summarize, the reliability of a mean depends *first* upon our having drawn an unbiased sample from the larger group or population which we are studying. When this condition has been met, and *only* then, the reliability of a mean is measured mathematically by its standard error which is based upon N (the number of cases) and the σ of the distribution. The formula for the standard error of the mean is

* Yule, G. U., *An Introduction to the Theory of Statistics* (10th ed., 1932), p. 257. For results of experiment, see Thorndike, E. L., *Empirical Studies in the Theory of Measurement*, Archives of Psychology, 3 (1907), 1–13.

$$SE_{\text{mean}} \text{ or } \sigma_M = \frac{\sigma}{\sqrt{N}} \tag{22}$$

(the standard error of the arithmetic mean when N *is large)**

This is an important and much-used formula. The standard error of the mean measures the extent to which this statistic is affected by errors of measurement (p. 398) as well as by differences which arise by chance from sample to sample. A decrease in σ or an increase in N will cause the standard error to become smaller numerically. A decrease in σ_M means that the amount by which the obtained mean probably misses the mean of the population is just so much less. In short, the reliability of an obtained mean *increases* as σ_M *decreases.*

A problem will illustrate the use and interpretation of formula (22).

> *Example (1)*† In 1883, the Anthropometric Committee of the British Association found the mean height of 8585 adult males in the British Isles to be 67.46 inches, with a σ of 2.57 inches. How reliable is this mean? How much does it probably diverge from the mean which would have been obtained had *all* adult males in the British Isles been measured?

We cannot answer these questions precisely when the value of the true mean is unknown (as here). But we can give a satisfactory answer provided we are willing to be in error once in 100 trials, or five times in 100 trials, or provided we are willing to take some other stipulated risk. Statisticians usually state the risk of error which they are willing to assume in a given investigation and their degree of confidence depends upon the "chances" they are willing to take (p. 187).

We know that our sample mean is 67.46 inches. Hence it is certain that 67.46 is *one* of the possible values that might arise through a random sampling of the given population. But

* Any N of 30 or more — 50 to be conservative — may be considered "large."

† Yule, G. U., *An Introduction to the Theory of Statistics* (10th ed. 1932), pp. 88–89, 112 and 141.

other values could also have arisen, and from a knowledge of sampling theory we can predict the probable range within which *all* of these possible sample means will lie. If we are willing to take the risk of being wrong five times in 100 trials, we can put the lowest mean obtainable from a sample at $67.46 - 1.96\sigma_M$, and the highest mean obtainable at $67.46 + 1.96\sigma_M$. If we are willing to take the risk of being wrong only *once* in 100 trials, we must put the lowest mean obtainable from a sample at $67.46 - 2.58\sigma_M$ and the highest mean obtainable at $67.46 + 2.58\sigma_M$. The reason for these limits ($\pm 1.96\sigma_M$ and $\pm 2.58\sigma_M$) is that sampling fluctuations around the population mean are known to follow the normal probability curve when samples are random. From Table 17, page 115, we find that 95% of the cases in a normal distribution fall between the limits $\pm 1.96\sigma_M$ (5% lying outside these limits); and 99% of the cases fall between the limits $\pm 2.58\sigma_M$ (1% lying outside these limits).

Now applying formula (22) we find the standard error of the mean, σ_M, to be $\frac{2.57}{\sqrt{8585}}$ or .028 inch. We can be confident, therefore, to the extent of risking a wrong answer five times in 100 trials that the range of sample means lies between $67.46 \pm 1.96 \times .028$, or $67.46 \pm .05$. The range of sample means from lowest to highest is therefore from 67.41 to 67.51.

The reliability of our mean depends upon the fact that we are quite confident that the true mean lies *somewhere* within this relatively narrow range. But our confidence does not amount to certainty since the given result depends upon our willingness to go wrong five times in 100 trials. If we wish to take a lesser risk (are willing to go wrong only *once* in 100 trials) we may conclude with greater confidence than before that the true mean lies within the range $67.46 \pm 2.58\sigma_M$ or between $67.46 - .07$ and $67.46 + .07$. Since the range within which the population parameter (true mean) probably falls is quite narrow in either case, we conclude that our obtained mean cannot be very far "off" from the true value, and that considerable confidence may be placed in its adequacy.

How the standard error measures the reliability or stability of an obtained mean may be more clearly shown perhaps in the following way: Suppose that we have calculated the mean height of each of 100 groups of men; that each group contains 8585 subjects; and that the groups or samples are drawn at random from the general population. The 100 means obtained from these samples will tend to differ slightly from one another owing to errors of sampling, or sampling fluctuations. Hence, not all samples will represent with equal fidelity the population from which they have been drawn. It can be shown mathematically that the frequency distribution of these sample means will fall into a normal distribution around the "true" or population mean as their measure of central tendency. Even when the *samples* are not normally distributed themselves, the *means* from such samples will tend toward a normal distribution. This "sampling distribution" of means measures the "errors" of sampling or fluctuations in mean values from sample to sample. In this hypothetical normal distribution of means we find relatively few *large* plus or minus deviations; and many *small* plus, *small* minus, and *zero* deviations. In short, the obtained means will hit very near to the true mean, or fairly close to it, more often than they will miss it by large amounts.

The mean of our distribution of 100 means is the best estimate of the "true" or population mean. And our *best estimate* of the σ of this distribution of means is the standard error of the mean which we have calculated. In other words, σ_M measures the spread of sample means around the true or population mean. It is because of this fact that the standard error of the mean becomes a measure of the amount by which *any* obtained mean *probably diverges* from the population mean

The results of our hypothetical experiment are represented graphically in Figure 42, page 187. The 100 sample means are represented by a normal frequency distribution around the *TM* (true mean) and σ_M is put equal to .028. The heights of the different ordinates (y's) represent the frequency of the various sample means. That the true mean is the most frequently obtained measure is shown by the fact that the ordinate at the *TM* is the maximum ordinate. The σ of a normal distribution when measured off in the plus and minus directions from the mean includes the middle 68.26% of the cases. About 68 of our 100 obtained means, therefore, may be expected to miss the *TM* by not *more* than $\pm$ 1 σ_M ($\pm$ .028 inch); and about 96 of our obtained means may be expected to miss the *TM* by $\pm$ 2 σ_M ($\pm$ .056 inch). Since our mean of 67.46 inches is *one* of these obtained means the probability is approximately .96 that 67.46 inches does not miss the true mean by more than $\pm$ .056 inch.

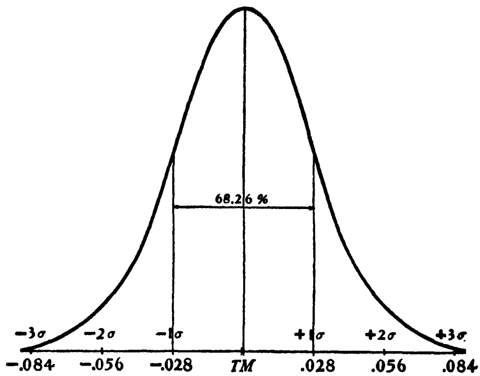


FIG. 42. Sample Distribution of Means Showing Variability of Obtained Means around the True or Population Mean (TM) in Terms of σ_M (.028).

(2) The *PE* of the Mean (PE_M)

The reliability of a mean may be determined by PE_M instead of by σ_M. PE_M is obtained by multiplying σ_M by .6745 (see p. 119). Thus

$$PE_M = \frac{.6745\sigma}{\sqrt{N}} \tag{23}$$

(*the probable error of the arithmetic mean when* N *is large, i.e., greater than 50*)

(3) Determining Limits of Accuracy

As we have seen, the reliability of an obtained mean will depend upon the likelihood of its having missed the true value by a large or small amount. An obvious difficulty in statements concerning reliability arises from our inability to say just how much the probable deviation of sample from population mean should be before it is to be judged "large." The sampling error allowable in a mean depends upon the purpose of the experiment, the standards of accuracy set up, the units in terms of which measurement is made and other factors.* An experi-

* Garrett, H. E., "Mean Differences and Individual Differences," *Human Biology*, 15 (1943), 155–170.

menter can never state categorically that his computed mean is — or is not — "reliable." But he can set up definite limits within which he may be quite confident of his result. Degree of confidence will depend upon the limits imposed. Fisher has proposed two accuracy limits, called respectively the .05 and the .01 levels (p. 201), and these may be accepted as standard for most experimental work.* We know from Table 17 that 95% of the cases in a normal distribution lie within the limits $\pm 1.96\sigma_M$. Hence, the odds are 95:5 or 19:1 that *any* sample mean will lie within these limits. Furthermore, since 99% of the cases in a normal distribution lie within the limits $\pm 2.58\sigma_M$, the odds are 99:1 that *any* sample mean will not differ from the population mean by more than $\pm 2.58\sigma_M$. In our height problem on page 184, we were able to say with considerable confidence (the odds are 19:1) that 67.46 inches does not differ from the true mean height by more than ± .05 inch. And we could say with still greater confidence (the odds are 99:1), that 67.46 inches does not differ from the true mean by more than ± .07 inch. The two limits, .05 and .01, mark off or define confidence intervals, the .01 level deserving greater respect than the .05 level.

(4) The *SE* of the Mean in Small Samples

Modern writers on statistics make a distinction between the standard deviation of the population and the standard deviation of a sample drawn from this population, often designating the population *SD* by σ, and the sample *SD* by s. It can be shown mathematically† that the sample *SD* systematically underestimates (is smaller than) the population σ, and this underestimation is more severe when samples are small. To correct this tendency toward negative bias, we should compute the standard deviation of a small sample by the formula

* Fisher, R. A., *The Design of Experiments* (1935), pp. 15–16, 38–43. Tippett, L. H. C., *The Methods of Statistics* (1937), pp. 69–71.

† Lindquist, E. F., *Statistical Analysis in Educational Research* (1940), pp. 48–50. For an interesting demonstration that s is our best estimate of the population σ, see Goulden, C. H., *Methods of Statistical Analysis* (1939), pp. 33–37.

$s = \sqrt{\frac{\Sigma x^2}{(N-1)}}$ rather than by the usual formula, $\sigma = \sqrt{\frac{\Sigma x^2}{N}}$ (p. 58). When N is large the correction effected by using $(N - 1)$ instead of N is negligible, but when N is small the correction may be considerable.

In the formula for the *SE* of the mean $\left(\sigma_M = \frac{\sigma}{\sqrt{N}}, \text{ p. } 184\right)$, the σ in the numerator is the population and *not* the sample *SD*. We never *actually* have the population σ; but we can estimate it, and our *best estimate* is s. When N is less than 50 or so, the formula for the *SE* of the mean should read:

$$\sigma_M = \frac{s}{\sqrt{N}} \tag{24}$$

where $$s = \sqrt{\frac{\Sigma x^2}{(N-1)}}$$

If the *SD* has already been computed by the formula $\sigma = \sqrt{\frac{\Sigma x^2}{N}}$, we can make the same correction in σ_M given in (24) by using the formula

$$\sigma_M = \frac{\sigma}{\sqrt{(N-1)}} \tag{24a}$$

(standard error of the mean in small samples,
i.e., N *less than 50)*

No matter what the size of N, formulas (24) and (24a) give the best estimate of the standard error of the mean, i.e., of the *SD* of the sampling distribution of means (Fig. 42, p. 187). In very large samples the correction effected by using (24) or (24a) is so small that formula (22) may be safely employed. But when N is less than 50 it is advisable to use the more exact formulas, and it is imperative when N is quite small — less than 10, say.

In small samples, the normal curve no longer tells us accurately the probability of a divergence of our sample mean

TABLE 29

Table of t

For Use in Determining the Reliability of Statistics.
If N is Large, Tables 17 and 18 May Be Used.

Example: An $(N - 1) = 35$ and $t = 2.03$ means that 5 times in 100 trials a divergence as large as that obtained may be expected *in the positive and negative directions.*

Degrees of Freedom	PROBABILITY (P)				
$(N - 1)$	0.50	0.10	0.05	0.02	0.01
1	$t = 1.000$	$t = 6.34$	$t = 12.71$	$t = 31.82$	$t = 63.66$
2	0.816	2.92	4.30	6.96	9.92
3	.765	2.35	3.18	4.54	5.84
4	.741	2.13	2.78	3.75	4.60
5	.727	2.02	2.57	3.36	4.03
6	.718	1.94	2.45	3.14	3.71
7	.711	1.90	2.36	3.00	3.50
8	.706	1.86	2.31	2.90	3.36
9	.703	1.83	2.26	2.82	3.25
10	.700	1.81	2.23	2.76	3.17
11	.697	1.80	2.20	2.72	3.11
12	.695	1.78	2.18	2.68	3.06
13	.694	1.77	2.16	2.65	3.01
14	.692	1.76	2.14	2.62	2.98
15	.691	1.75	2.13	2.60	2.95
16	.690	1.75	2.12	2.58	2.92
17	.689	1.74	2.11	2.57	2.90
18	.688	1.73	2.10	2.55	2.88
19	.688	1.73	2.09	2.54	2.86
20	.687	1.72	2.09	2.53	2.84
21	.686	1.72	2.08	2.52	2.83
22	.686	1.72	2.07	2.51	2.82
23	.685	1.71	2.07	2.50	2.81
24	.685	1.71	2.06	2.49	2.80
25	.684	1.71	2.06	2.48	2.79
26	.684	1.71	2.06	2.48	2.78
27	.684	1.70	2.05	2.47	2.77
28	.683	1.70	2.05	2.47	2.76
29	.683	1.70	2.04	2.46	2.76
30	.683	1.70	2.04	2.46	2.75
35	.682	1.69	2.03	2.44	2.72
40	.681	1.68	2.02	2.42	2.71
45	.680	1.68	2.02	2.41	2.69
50	.679	1.68	2.01	2.40	2.68
60	.678	1.67	2.00	2.39	2.66
70	.678	1.67	2.00	2.38	2.65
80	.677	1.66	1.99	2.38	2.64
90	.677	1.66	1.99	2.37	2.63

Degrees of Freedom ($N-1$)	PROBABILITY (P) 0.50	0.10	0.05	0.02	0.01
100	.677	1.66	1.98	2.36	2.63
125	.676	1.66	1.98	2.36	2.62
150	.676	1.66	1.98	2.35	2.61
200	.675	1.65	1.97	2.35	2.60
300	.675	1.65	1.97	2'34	2.59
400	.675	1.65	1.97	2.34	2.59
500	.674	1.65	1.96	2.33	2.59
1000	.674	1.65	1.96	2.33	2.58
∞	.674	1.65	1.96	2.33	2.58

from the population mean. The sampling distribution to be used when N is small is not strictly normal; its "shoulders" are higher than in the normal curve and the probability of extreme deviations somewhat greater. Selected values for this sampling distribution, called "Student's" distribution,* are given in Table 29. For N's differing in size, this table gives the $\pm$ t-distances beyond which (i.e., to left *and* right) certain percentages of "Student's" distribution lie (.50, .10, .05, .02, .01). We may illustrate the use of Table 29 in small samples with a problem.

> *Example* (*2*) Ten measures of reaction time to a light stimulus are taken from one practiced observer. The mean is 175.50ms, and the σ is 5.82ms. Determine the .05 and the .01 limits of accuracy for this mean.

From formula (24a) $\sigma_M = \dfrac{5.82}{\sqrt{10-1}} = \dfrac{5.82}{3} = 1.94$ms. Ten observations have 9 "degrees of freedom,"† and from Table 29, for 9 or $(N - 1)$ degrees of freedom, we read that $t = 2.26$ (at the .05 level), and $t = 3.25$ (at the .01 level). *The quantity t is distance from the mean expressed in terms of the standard error of*

* Fisher, R. A., *Statistical Methods for Research Workers* (8th ed., 1941), pp. 116–117.

† When the sum (or mean) of 10 measures is known, only 9 may be selected "freely," as the sum (or M) fixes the 10th. Accordingly, there are 9 degrees of freedom for 10 measures and in general $(N-1)$ degrees of freedom for N measures. See also page 257.

the mean (i.e., $t = \pm x/\sigma_M$). From the first t we know that [when $(N - 1) = 9$] 95% of the sampling distribution fall between the mean and $\pm 2.26\sigma_M$ and 5% fall outside of these limits. From the second t we know that 99% of the sampling distribution fall between the mean and $\pm 3.25\sigma_M$, and 1% falls outside these limits. The probability is .95, therefore, that our sample mean of 175.50ms does *not* diverge from the population mean by more than $\pm$ 4.38ms ($\pm 2.26 \times 1.94$); and the probability is .05 that its divergence is greater than $\pm$ 4.38ms. At the .01 level, the probability is .99 that our mean of 175.50ms does *not* diverge from the population mean by more than $\pm$ 6.31ms ($\pm 3.25 \times 1.94$); and the probability is .01 that its divergence is greater than $\pm$ 6.31ms.

Several points in the solution of this problem deserve comment as they illustrate clearly the difference between the treatment of large and small samples. In the first place, had we used formula (22) instead of the correct formula (24a), the *SE* of the mean would have been 1.75 instead of 1.94; i.e., 10% too small. Again, the .05 and .01 accuracy limits in the normal curve are, as we have seen, $\pm 1.96\sigma_M$ and $\pm 2.58\sigma_M$, respectively. These limits are 15% and 20% smaller than the corresponding t limits $\pm$ 2.26 and $\pm$ 3.25 got from Table 29 when $(N - 1)$ is 9. It is clear, therefore, that when N is small, use of formula (22) will cause a calculated mean to appear *more* reliable than it actually is.

The reader should note that if formula (24a) and Table 29 are used in determining the reliability of the mean in our height problem (p. 184), results will not differ to the second decimal from those got with formula (22) and Table 17. This is because of the very large sample (8585) there used. As N increases, entries in Table 29 approach more and more closely the corresponding normal curve entries in Table 17. In the normal curve, for instance (Table 17), 10% of the distribution lie beyond the limits $\pm 1.65\sigma_M$, 5% beyond the limits $\pm 1.96\sigma_M$ and 1% beyond $\pm 2.58\sigma_M$. In Table 29 the corresponding t limits for $(N - 1) = 50$, are $\pm$ 1.68, $\pm$ 2.01, $\pm$ 2.68; for $(N - 1) = 100$,

the limits are ± 1.66, ± 1.98, ± 2.63. When N is *very* large (see last entries in Table 29) the points beyond which specified percents of the distribution lie are virtually the same in Table 29 as in Table 17, and "Student's" distribution becomes a normal probability curve. Table 29 may be generally used, then, with large as well as with small samples.

2. The Reliability of the Median

The standard error and the probable error of the median may be computed directly from formulas for determining the reliability of the mean. The σ_{Mdn} and PE_{Mdn} are 1.2533 (roughly 5/4) times the σ_M and the PE_M, respectively. Thus

$$\sigma_{Mdn} = \frac{1.2533\sigma}{\sqrt{N}} \tag{25}$$

(*standard error of the median when* N *is large*)

$$PE_{Mdn} = \frac{1.2533 \times .6745\sigma}{\sqrt{N}} = \frac{.8454\sigma}{\sqrt{N}} \tag{26}$$

or

$$PE_{Mdn} = \frac{1.2533Q^*}{\sqrt{N}} \tag{26a}$$

(*probable error of the median when* N *is large*)

When samples are small (less than 50, say), $(N - 1)$ should replace N in the denominators of these formulas, and Table 29 should be used in setting up accuracy limits at different levels of confidence.

An example will illustrate the use of formula (26a).

Example (3) On the Trabue Language Scale A,† 801 twelve-year-old boys made the following record: Median = 21.4; Q = 4.9. How reliable is this median? How well does it represent the median of twelve-year-old boys in general on the given scale?

* The quartile deviation calculated from a frequency distribution is Q, not PE.

† Trabue, M. R., *Completion Test Language Scales*, Teachers College, Columbia University Contributions to Education, 77 (1916), 15.

Since N is quite large, we may use formula (26a) to find PE_{Mdn} equal to .22. In a normal distribution, the middle 95% of cases lie between the mean and $\pm$ 2.9PE, and the middle 99% between the median and $\pm$ 3.8PE (see Table 18). We may say with considerable confidence, therefore (odds, 19:1), that 21.4 does not diverge from its true value by more than $\pm$ .64($\pm$ 2.9 $\times$ .22); and with much greater confidence (odds, 99:1) that 21.4 does not miss the population median by more than $\pm$ .84($\pm$ 3.8 $\times$ .22).

III. The Reliability of Measures of Variability

1. The Reliability of the Standard Deviation, or σ

As was true of the mean and median, the reliability of an obtained standard deviation is determined by calculating the probable discrepancy between the sample σ and the true σ. A true σ is the standard deviation of the population from which the sample was drawn. The formula for calculating the reliability of an obtained σ is:

$$SE_\sigma \text{ or } \sigma_\sigma = \frac{\sigma}{\sqrt{2N}} \tag{27}$$

(*standard error of a standard deviation when* N *is large*)

When N is less than about 50, formula (27) should read:

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2(N-1)}} \text{ or } \frac{s}{\sqrt{2N}} \tag{27a}$$

(*standard error of a standard deviation when* N *is small*)

On page 184 we found that for 8585 British males, the standard deviation around the mean of 67.46 inches was 2.57 inches. Since the sample is large, we may use formula (27) to find $\sigma_\sigma = \frac{2.57}{\sqrt{2 \times 8585}} = .02$ inch. We may be confident (probability is .95) that the population σ is not larger than 2.61 nor smaller than 2.53 inches (2.57 $\pm$ 1.96 $\times$ .02). And we may feel very confident (probability .99) that the population σ is not greater than 2.62 nor less than 2.52 (2.57 $\pm$ 2.58 $\times$ .02). These

relations are shown in Figure 43 in which the true σ is represented as the mean of a sampling distribution of σ's, i.e., the distribution of σ's computed from successive samples. The standard deviation of this normal distribution is .02, the standard error of σ.

In the problem on page 191 the calculated *SD* was 5.82ms, around the mean of 175.50ms. Since N is only 10, we use formula (27a) to get $\sigma_\sigma = \frac{5.82}{\sqrt{2(10-1)}} = 1.37$ms. From Table

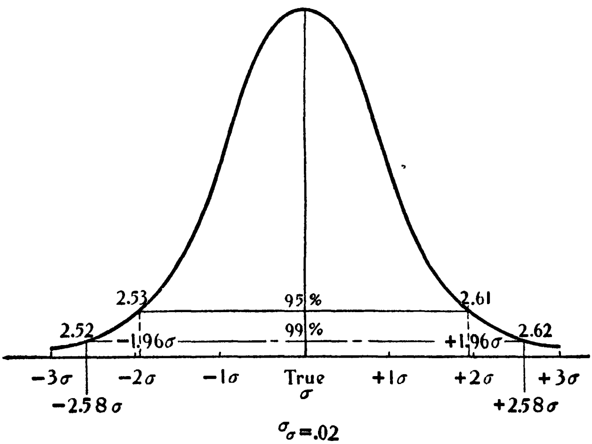


FIG. 43.

29 we find, as before, that for $(N - 1) = 9$ the accuracy limits at .05 and .01 are $\pm$ 2.26 and $\pm$ 3.25, respectively. We may feel confident, therefore (the probability is .95), that the population σ is not larger than 8.92ms nor smaller than 2.72ms $(5.82 \pm 2.26 \times 1.37)$. And we may feel very confident (the probability is .99) that the population σ is not larger than 10.27ms nor smaller than 1.37ms $(5.82 \pm 3.25 \times 1.37)$. Note that if formula (27) had been used, the standard error of our σ would have been $\frac{5.82}{\sqrt{20}}$ or 1.30ms instead of 1.37ms. Thus had we used large instead of small sample methods, the limits of

accuracy at the .05 level would have been 8.37ms and 3.27ms ($5.82 \pm 1.96 \times 1.30$) instead of 8.92ms and 2.72ms, the correct values.

The reader should note two facts: (1) the relatively wide range of likely values for σ (high unreliability) when N is small; and (2) the greater *apparent* reliability of the standard deviation when large sample methods are incorrectly used with small samples. Because of (2), it is wise to use formula (27a) and Table 29 even when N is fairly large.

2. The Reliability of the Quartile Deviation or Q

The reliability of the Q of a distribution may be found from the formula

$$\sigma_Q = \frac{1.11\sigma}{\sqrt{2N}} \tag{28}$$

(*standard error of* Q *in terms of the* σ *of the distribution*)*

or from the formula

$$\sigma_Q = \frac{1.65Q}{\sqrt{2N}} \tag{28a}$$

(*standard error of* Q *in terms of the* Q *of the distribution*)*

On page 193, the median score of the 801 twelve-year-old boys who took the Trabue Completion Test, Scale A, was 21.4 with a Q of 4.9. Since N is large, we may use formula (28a) to find σ_Q equal to .20. Adopting the .05 level, we may be confident that the population Q lies between 5.3 and 4.5 ($4.9 \pm 1.96 \times .20$). Stated differently, there is only 1 chance in 20 that the sample Q of 4.9 differs from the population Q by more than $\pm$ .40 (i.e., $\pm 1.96 \times .20$).

* When N is less than 50, formulas (28) and (28a) should be written $\sigma_Q = \frac{1.11\sigma}{\sqrt{2(N-1)}}$ or $\frac{1.11s}{\sqrt{2N}}$ and $\sigma_Q = \frac{1.65Q}{\sqrt{2(N-1)}}$ and Table 29 used in tests of significance.

IV. The Reliability of the Difference between Two Measures

1. The Reliability of the Difference between Two Means

Suppose we wish to discover whether there is any difference between fifth-grade boys and fifth-grade girls in their knowledge of words. The usual method of attacking this problem is to select a large and random sample of fifth-grade boys and girls; administer a vocabulary test; compute the means; and find the difference between the two means. If this difference is five points, let us say, in favor of the girls, this result — on the face of it — may be taken as evidence that the typical fifth-grade girl knows more words than the typical fifth-grade boy. We cannot be certain of this conclusion, however, if all we have is the obtained difference of five points, as it is quite possible that the difference between the means of other samples of boys and girls (comparable to our own groups) might turn out to be zero or might even be reversed in favor of the boys.

When *can* we feel reasonably sure that a difference is "real" and not accidental? The answer to this question can never be absolute; it always involves a statement of probability and is usually expressed in terms of the accuracy limits discussed in the last section. A difference is said to be *significant* (i.e., reliable or dependable) when the evidence is strong that the result found cannot be attributed solely to accidents of sampling. By the same token, a difference is *non-significant* when we are confident that it might easily have arisen from sampling fluctuations — and hence implies no "real" difference.

Clearly it is important that we have some way of estimating the significance of an obtained difference; that is, some way of telling whether two groups are *sufficiently* different to enable us to say with confidence that no matter how often other similarly selected samples are compared, some difference will persist. Furthermore, and equally important, if the obtained difference is *not* significant, we want to know, if possible, how near it approaches to significance.

(1) The Standard Error of the Difference When Means Are Uncorrelated (σ_D)

The formula for calculating the significance of the difference between two sample means when we are dealing with independent or uncorrelated measures is

$$\sigma_D \text{ or } \sigma_{M_1-M_2} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2}} \tag{29}$$

(*standard error of the difference between two uncorrelated means,* N's *large*)*

in which σ_{M_1} is the standard error of the mean of the first group; σ_{M_2} is the standard error of the mean of the second group; and σ_D is the standard error of the difference between the two means. Means are *uncorrelated* when calculated from different groups, or from uncorrelated tests administered to the same group. From formula (29) it is clear that, to find the reliability of the difference between two means, we must first know the reliabilities of the means themselves.

The application and interpretation of formula (29) may be illustrated by the following example:

Example (1) In a study of the intelligence of the foreign-born white draft during World War I, a sample of 611 native-born Norwegians and a sample of 129 native-born Belgians were found to test as follows upon the "combined scale"†:

Country of Birth	Number of Cases	Mean Score	σ
Norway	611	12.98	2.47
Belgium	129	12.79	2.42

The difference between the two obtained means is .19 (12.98 − 12.79) in favor of the Norwegians. Is this difference significant? That is to say, would further testing of similar samples

* When the *PE*'s of the means have been computed, the *PE* of the difference between two means is

$$PE_D \text{ or } PE_{M_1-M_2} = \sqrt{PE^2_{M_1} + PE^2_{M_2}} \tag{30}$$

† The "combined scale" included the eight Alpha tests, the Stanford-Binet, and tests 4, 5, 6, and 7 from Beta. The maximum score was 25. For the data given in this problem, see Brigham, C. C., *A Study of American Intelligence* (1923), pp. 120–121.

of Norwegians and Belgians give virtually the same result; or is it probable that the mean difference would be reduced to zero, or even reversed in favor of the Belgians? To answer these questions we must first compute the standard errors of the means of Norwegians and Belgians, and from these data find the reliability of the difference between the means. By formula (22), the standard errors of the two means are

$$\text{Norwegians:} \quad \sigma_{M_1} = \frac{2.47}{\sqrt{611}} = .0999$$

$$\text{Belgians:} \quad \sigma_{M_2} = \frac{2.42}{\sqrt{129}} = .2130$$

Substituting these standard errors in formula (29), we have

$$\sigma_D = \sqrt{(.0999)^2 + (.2130)^2} = .24 \text{ (to two decimals)}$$

The actual difference between the means of Norwegians and Belgians, then, is .19, and the *SE* of this difference (σ_D) is .24. Let us assume that the difference between the *population means* of Norwegians and Belgians is zero, and that except for accidental errors mean differences from sample to sample would *all* be zero. In making this assumption we are setting up the "null hypothesis," a proposition somewhat akin to the legal principle that a man is innocent until he is proved guilty (p. 232). In our problem, for example, we inquire whether — in view of its *SE* — the mean difference of .19 is really large enough to cast grave doubt upon (i.e., disprove) the null hypothesis.

As a first step in testing our hypothesis, we compute a "critical ratio" or *CR*, found by dividing the obtained difference by its *SE* ($D/\sigma_D = CR$). In the present problem, the $CR = .19/.24$, or .79. The sampling distribution of differences among sample means is known to be normal when N is reasonably large. Hence we may set up a normal distribution like that shown in Figure 44 in which the mean is set at zero ("true difference") and the σ of the distribution of differences is σ_D, or .24.* The *CR* tells us that our obtained difference of .19 falls at a point $.79\sigma_D$ from

* σ_D is the best estimate we have of the *SD* of the sampling distribution of differences (p. 189).

the hypothetical mean of zero; and a difference of $-.19$ will, of course, fall at $-.79\sigma_D$. The value $-.19$ is obtained when the mean of the Belgians is higher than the mean of the Norwegians by .19.

Now from Table 17 we know that $29\% \times 2$, or 58% of the cases in a normal distribution fall between the mean and $\pm .79\sigma_D$; and 42% of the cases fall outside of these limits. Even when the true difference is zero, then, we can expect differences *larger* than $\pm .19$ to occur *by chance* forty-two times in 100 comparisons of Norwegians and Belgians. A difference of $\pm .19$, therefore, might easily arise from sampling errors and is clearly *not* significant. Accordingly, we *retain* the null hypothesis and conclude with confidence that — on present evidence — there is no real difference between Norwegians and Belgians on the combined scale. When the null hypothesis is not disproved (as here) the result is often stated as follows: there is good reason to believe that our two samples were drawn from the *same* parent population and differ only by sampling errors.

So far we have dealt with the probability that, on the null hypothesis, the Norwegians are better than the Belgians by .19,

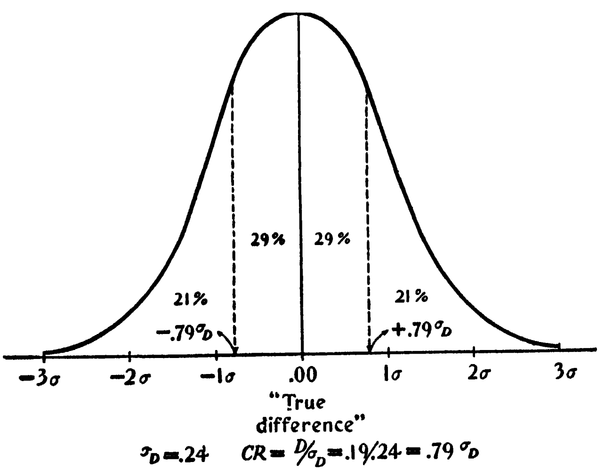


FIG. 44.

and the probability that the Belgians are better than the Norwegians by .19 (— .19). In many, perhaps most, experiments, however, we are mainly concerned with the *direction* of the difference, i.e., with the probability that the obtained difference or a larger one might have arisen on the null hypothesis. In studying the effects of practice and other experimental factors, for instance, usually we want to know the probability that one group (the experimental, say) is "really" better than the other group (the control); or we inquire the probability that boys are better than girls in mechanical aptitude or in some other ability. In such cases we deal only with the *positive* end of the sampling distribution of differences. To illustrate, we have found in Example (1) that the Norwegians are .19 point higher than the Belgians. What is the probability that on the average the Norwegians will always score higher than the Belgians by .19 or more? From Figure 44 we know that a difference of .19 will be exceeded by chance 21% of the time. Even when the true difference is zero, then, we could expect to find the Norwegians better than the Belgians by more than .19 point in 1/5 of our comparisons. The difference of .19 (or more) might readily be ascribed to chance, therefore (its probability $P = .21$), and there is no reason for believing the Norwegians to be better in general than the Belgians on the combined scale.

(2) Interpretation of Differences in Terms of Significance Levels

(*a*) The .05 level of significance

An investigator often sets up some arbitrary standard of significance on the basis of which he rejects or retains the null hypothesis. From Table 17 or 29 (last line in Table) we find that 1.96 marks the point in the normal distribution to the left and right of which lie 5% of the cases ($2\frac{1}{2}$% at each end). If a *CR* is 1.96 and the *N* is large, therefore, we reject the null hypothesis with some confidence on the grounds that the given difference can hardly be attributed to sampling errors.

The *CR* of .79 in the problem of Norwegians and Belgians falls far below the .05 level of significance, for which a *CR* of 1.96 is necessary. All we need say in this problem, therefore,

is that we retain the null hypothesis with confidence since on the evidence there is no reason to suspect a true mean difference between Norwegians and Belgians.

Significance levels may also be used when we are interested in the probability that one group is better than the other. From Table 29 we know that 10% (P) of the cases in a normal distribution lie to the *left* and *right* of $t = 1.65$; hence, 5% ($P/2$) lie to the *right* of 1.65. If a CR is 1.65, therefore, we can say with confidence that (on the assumption of a true difference of zero) only once in twenty trials would a larger positive difference than that obtained appear by chance.

From Figure 44 we have found that twenty-one times in 100 trials a difference between Norwegians and Belgians of more than + .19 might be expected on the null hypothesis. Because of the large chance expectation of a positive difference of .19 or more, we can feel sure that the Norwegians are *not* superior to the Belgians on the combined scale.

A second example may serve to clarify certain points discussed above. Suppose that the difference between the means of an experimental Group A and a control Group B upon Test X is six points, that σ_D is 3, and N's are quite large. Since the CR of 6/3 or 2 is slightly greater than 1.96, this result may be considered significant at the .05 level. We reject the null hypothesis with confidence, therefore, since it is quite unlikely (odds 19:1) that a critical ratio of 2 (absolute difference of $\pm$ 6) would occur if the difference between the population means of A and B were in fact zero. We could expect a difference of more than 6 (*positive* direction) to appear in favor of the experimental group not more than two or three times in 100 trials. Hence, we are justified in asserting that Group A is, in general, superior to Group B in Test X. Figure 45 shows graphically the relations represented in this problem.

Still another way of interpreting the significance of a difference is in terms of the "accuracy limits" discussed on page 187. In the problem of Norwegians and Belgians, for instance, we obtained a difference of .19 with a σ_D of .24. We may be confident,

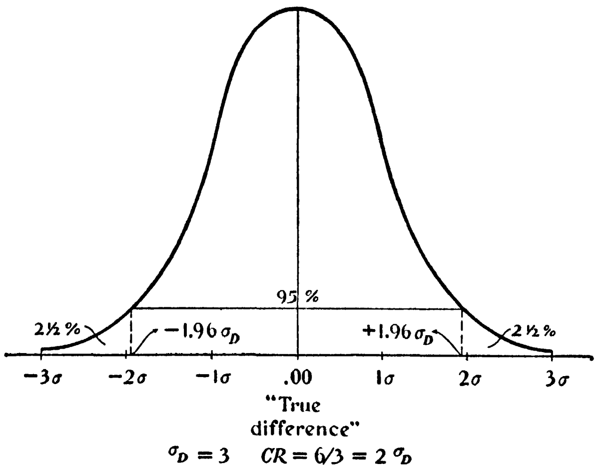


FIG. 45.

therefore (odds 19:1), that the difference between Norwegians and Belgians lies within the limits − .28 and + .66 (.19 ± 1.96 × .24). Since the lower limit of this range is negative it is quite clear (as found before) that the difference between these groups could well be zero. In the second problem, the difference between control and experimental groups was six points with a σ_D of 3. This difference is twice its standard error and hence is significant (p. 208). We not only assert a significant difference, therefore, but put its value with considerable confidence as lying between 0 and 12 points (6 ± 1.96 × 3).

(*b*) The .01 level of significance

While the .05 level is sufficiently exacting for most investigations, the .01 level is demanded by many research workers. From Table 17 or Table 29 (last line) we read that ± 2.58 mark the points in the normal curve to the left and right of which lies 1% of the cases. If a *CR* is 2.58 or more, therefore, and *N*'s are large, we reject the null hypothesis with great confidence as only once in 100 trials would a larger difference arise from sampling errors, when the true difference is zero. If the critical ratio is 2.33 (P = .02 and $P/2$ = .01), we may be very

confident (odds 99:1) that the group now ahead is really superior to the second group in mean attainment.

(3) The Reliability of the Difference between Means in Small Independent Samples

When the N's of two independent samples are small (less than 50, say), the SE's of the means should be calculated by formula (24a) or some variation of it. Table 29 may be used conveniently in testing the significance of the critical ratio or t. An example will demonstrate the method to be employed.

Example (2) A test of mechanical aptitude is administered to six boys in Class 1, and to ten boys in Class 2 of a given vocational school. Is Class 1 significantly better than Class 2? Data are as follows:

Class 1 Scores	x	x^2	Class 2 Scores	x	x^2
28	− 2	4	30	2	4
35	5	25	26	− 2	4
32	2	4	25	− 3	9
24	− 6	36	34	6	36
26	− 4	16	20	− 8	64
35	5	25	28	0	0
6)180		110	31	3	9
30(M_1)			24	− 4	16
			32	4	16
			30	2	4
			10)280		162
			28(M_2)		

$(N_1 - 1) = 5$

$(N_2 - 1) = \frac{9}{14}$

$$SD \text{ or } s = \sqrt{\frac{162 + 110}{14}} = 4.41 \quad \text{by (31)}$$

$$SE_D = 4.41\sqrt{\frac{16}{60}} = 2.28 \quad \text{by (32)}$$

$$t = \frac{30 - 28}{2.28} = .88$$

For $(N_1 - 1) + (N_2 - 1)$ or 14 degrees of freedom, the .10 level for t is 1.76 (Table 29).

The mean of the six boys in Class 1 is 30, the mean of the ten boys in Class 2 is 28 and the mean difference of 2 is to be tested for significance. When two samples are small (as here) we get a better estimate of the population *SD* by pooling the sums of squares from the two groups and computing one *SD*. The justification for this pooling procedure is that on the null hypothesis the real difference between the two classes is zero; hence the two samples may be treated as though they were drawn from the same population.* Moreover, increasing the N gives a more stable *SD* based on *all* of the observations. The sum of the squares in Class 1 around the mean of 30 is 110; and the sum of the squares in Class 2 around the mean of 28 is 162. The degrees of freedom in Class 1 are $(N_1 - 1)$ or 5, and the degrees of freedom in Class 2 are $(N_2 - 1)$ or 9. By formula (31), $s = \sqrt{\frac{110 + 162}{14}}$ or 4.41, and this *SD* serves as the standard deviation for each of our groups. The *SE* of M_1 is $\frac{4.41}{\sqrt{6}}$ and the *SE* of M_2 is $\frac{4.41}{\sqrt{10}}$. Combining these by formula (29) we have $SE_D = \sqrt{\frac{(4.41)^2}{6} + \frac{(4.41)^2}{10}} = 4.41\sqrt{\frac{16}{60}} = 2.28$. Formula (32), on page 206, combines the two SE_M's directly.

The *CR* or t is $\frac{D}{\sigma_D}$ or $2/2.28 = .88$. The *df*† in the two groups (viz. 5 and 9) are combined to give 14*df* to be used in evaluating the mean difference. From Table 29 for 14 degrees of freedom we find the entry 1.76 at the .10 level. The critical ratio of .88 falls far below 1.76. Hence the difference of + 2 is *not* significant at the .05 level and there is no reason to believe Class 1 superior to Class 2. It must be remembered that at .10, 5% of our t's (*CR*'s) lie to the *right* of + 1.76 and 5% lie to the *left* of − 1.76. The limit at .10 (not at .05) must be taken, therefore, to give the .05 significance level, if we are interested (as here) in knowing the probability that *the given difference or a*

* We assume the null hypothesis to hold until it is disproved.

† *df* = degrees of freedom.

greater positive one might arise from sampling errors. Figure 45a illustrates this point.

The formulas used in testing the significance of a mean difference in small independent samples may be written as follows:

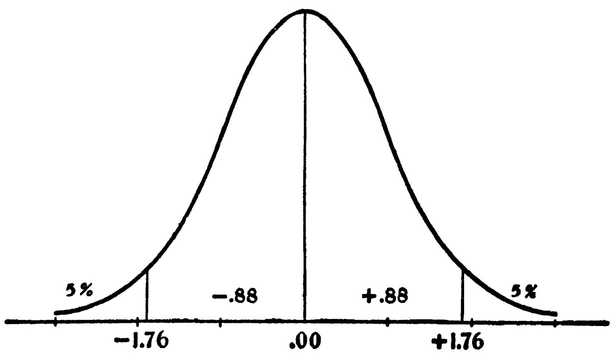
$$SD \text{ or } s = \sqrt{\frac{\Sigma(X_1 - M_1)^2 + \Sigma(X_2 - M_2)^2}{(N_1 - 1) + (N_2 - 1)}} \tag{31}$$

(standard deviation when two small independent samples are pooled)

$$SE_D \text{ or } s_D = s\sqrt{\frac{N_1 + N_2}{N_1 N_2}} \tag{32}$$

(standard error of the difference between means in small independent samples)

In formula (31), $\Sigma(X_1 - M_1)^2$ or Σx_1^2 is the sum of the squared deviations around the mean of sample 1; and $\Sigma(X_2 - M_2)^2$ or



For 14 degrees of freedom 5% of the distribution lie to the left and 5% to the right of 1.76*t*. (Table 29.)

FIG. 45a.

Σx_2^2 is the sum of the squared deviations around the mean of sample 2. These sums of squares are combined as shown above, in order to give a better estimate of the *SD*. In computing the *SE* of the difference between means, the *SE* of each mean is calculated from the same *SD*; hence formula (32) enables us to calculate SE_D directly.

A second example will serve to illustrate further the use of significance levels when samples are small.

Example (3) On an arithmetic reasoning test thirty-one ten-year-old boys and forty-two ten-year-old girls made the following scores:

	Mean	σ	N
Boys:	40.39	8.69	31
Girls:	35.81	8.33	42

Is the mean difference of 4.58 between boys and girls significant?

We may calculate the σ_D directly by formula (29) to be: —

$\sigma_D = \sqrt{\frac{(8.69)^2}{30} + \frac{(8.33)^2}{41}} = 2.05.$ Note that the standard errors of the means are calculated by formula (24a): $(N - 1)$ for boys is 30, and $(N - 1)$ for girls, 41.

The t or critical ratio is 4.58/2.05, or 2.23, and the degrees of freedom to be used in testing the significance of the difference (i.e., 4.58) are $30 + 41$, or 71. We may take the t's for 70 degrees of freedom in Table 29 without interpolation as these t's furnish a close approximation to the t's for 71 degrees of freedom. When the degrees of freedom equal 70, a t of $\pm$ 2.00 or more may be expected on the null hypothesis 5% of the time, and a t of $\pm$ 2.65 or more may be expected 1% of the time. The obtained t of 2.23 passes the .05 but not the .01 level. We may, therefore, reject the null hypothesis with considerable confidence. Moreover, we can assert not only that the difference between boys and girls is significant, but that its value (odds 19:1) lies between .48 and 8.68 ($4.58 \pm 2.00 \times 2.05$).

(4) The Use of Table 29 in Determining the Significance of a Difference

At the risk of repetition it may be helpful to summarize the applications of Table 29 to the problem of determining the reliability of differences. For varying degrees of freedom, Table 29 gives the values — t's or CR's — to the left and right of which lie certain proportions of "Student's" distribution (p. 190).

The t's from Table 29 equal the critical ratios $\left(\frac{D}{\sigma_D}\right)$ of the normal curve exactly when N is very large (i.e., ∞), and approximate them quite closely when N's are 50 or more.

In general, t's are tested against the null hypothesis, i.e., against the assumption that there is no true difference between the population means being compared, and that our two samples differ only through sampling accidents. Depending upon the evidence, we refute or retain the null hypothesis. When groups are independent, the degrees of freedom used in testing the significance of a difference equal $(N_1 - 1) + (N_2 - 1)$, where N_1 is the size of the first, and N_2 the size of the second sample. If the degrees of freedom equal 20, we reject the null hypothesis at the .05 level if t equals 2.09, and at the .01 level if t equals 2.84. For t's less than 2.09 we accept the null hypothesis and mark the difference "not significant." When the degrees of freedom equal 30, a t of .683 stands for a difference which might be expected to occur fifty times in 100 trials through sampling errors alone. For *any* P (probability) greater than .05, the null hypothesis is retained and the difference is marked not significant.

For many years it has been customary for investigators to demand a critical ratio of 3 or more before a difference is regarded as significant. This extremely high standard sets up a confidence level which is probably not warranted in many experimental studies.

(5) The Standard Error of the Difference between Two Means, When Means Are Correlated

(*a*) Single Group Method

The last sections have dealt with the problem of determining whether the difference between two means is significant when these means represent the performance of different groups — boys and girls, Norwegians and Belgians, and the like. A closely related problem is concerned with the significance of the difference between two means obtained from the *same* test administered to the *same* group upon different occasions. This

is called the "single group" method. Suppose, for example, that we have administered a test to a group of children and after two weeks have repeated the test. We wish to measure the effect of practice or of intervening training upon the final scores; or to estimate the effect of some activity interpolated between test and retest. In order to determine the significance of the difference between the means obtained in the initial and final testing, we must use the formula

$$\sigma_D = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2} - 2r_{12}\sigma_{M_1}\sigma_{M_2}} \tag{33}$$

(*standard error of the difference between correlated means*)

in which σ_{M_1} and σ_{M_2} are the standard errors of the initial and final test means, and r_{12} is the coefficient of correlation between scores made on the initial and final tests.* An illustration will bring out the difference between formula (29) and formula (33).

Example (4) At the beginning of the school year, the mean score of a group of sixty-five sixth-grade children upon an educational achievement test in reading was 45.00 with a σ of 6.00. At the end of the school year, the mean score on an equivalent form of the same test was 50.00 with a σ of 5.00. The correlation between scores made on the initial and final testing was .60. Has the class made significant progress in reading during the year?

We may tabulate our data as follows:

	Initial Test		Final Test
No. of children:	65		65
Mean score:	45.00 (M_1)		50.00 (M_2)
Standard deviations:	6.00 (σ_1)		5.00 (σ_2)
Standard error of the mean:	.75 (σ_{M_1})†		.63 (σ_{M_2})†
Difference between means:		5.00	
Correlation between initial and final tests:		.60	

* The correlation between the means of successive samples drawn from a given population equals the correlation between test scores, the means of which are being compared. See Kelley, T. L., *Statistical Method* (1923), p. 178.

† By formula (24a)

Substituting in formula (33), we get

$$\sigma_D = \sqrt{(.75)^2 + (.63)^2 - 2 \times .60 \times .75 \times .63} = .63$$

Since there are sixty-five children, there are sixty-five pairs of scores, and sixty-five differences. The number of degrees of freedom, accordingly, is 65 − 1 or 64. The critical ratio, t or D/σ_D, is 5.00/.63, or 7.9. The t for $N - 1 = 64$ is 2.39 at the .02 level (Table 29). As our t is much larger than 2.39 the probability is far less than .01 that the gain (p. 202) can be attributed to sampling errors. It is clear, therefore, that this class made significant progress in reading during the school year.

When groups are small, a method slightly different from that given above is to be preferred when we are evaluating the difference between two correlated means. An example will serve as an illustration:

Example (*5*) Twelve subjects are given five successive trials upon a symbol-digit learning test. Data for the first and the fifth trials are as follows:

	1st trial	5th trial	Diff. (5 − 1)
Means:	160.42	171.85	11.43
σ:			14.05

The mean gain is 11.43, and the *SD* around this mean is 14.05. Is the gain due to practice significant?

From formula (22) the *SE* of the mean gain $\left[\frac{14.05}{\sqrt{12}}\right]$ is 4.35.

On the null hypothesis (i.e., with respect to a mean gain of zero) we wish to test the significance of our gain of 11.43. The *CR* or t is 11.43/4.35 or 2.63. For 11 degrees of freedom $[(N - 1) = 11]$, we find from Table 29 that a t of 2.72 (column .02) will be exceeded in the *positive* direction in 1% of the trials. From the column headed .10, we find that a t of 1.80 will be exceeded in the *positive* direction in 5% of the trials. Our mean *gain* of 11.43 ($t = 2.63$) is significant at the .05 level, therefore, and almost significant at the .01 level. Note again that we take entries from the .10 and .02 columns (for significance levels .05 and .01).

when we are interested in the probability of a gain (positive difference) as large or larger than 11.43.

In problems like this, dealing with mean gain involves less calculation and is to be preferred to the method of calculating SE's for each mean, an SE of the difference, and the correlation between initial and final scores.

(*b*) Equivalent Groups Method

Formula (33) is often employed in experiments which make use of the method of *equivalent groups*. The equivalent groups method enables us to evaluate the effect of one or more experimentally varied conditions (experimental factors) as compared with the absence of these factors (control conditions). The following problem is typical of many to which the equivalent group technique is applicable:

> *Example* (*6*) Two groups, X and Y, of seventh-grade children are paired child for child for age and for score upon Form A of the Otis Group Intelligence Scale. Three weeks later, both groups are given Form B of the same test. Before the second test, Group X, the experimental group, is praised for its performance on the first test and urged to better its score if possible. Group Y, the control group, is given the second test without comment. Will the incentive (praise) serve to increase significantly the final score of Group X over Group Y?

The relevant data may be tabulated as follows:

TABLE 30

	Experimental Group X		Control Group Y	
No. of children in each group:	72		72	
Mean scores on Form A, initial test:	80.42		80.51	
SD on Form A, initial test:	23.61		23.46	
Mean scores on Form B, final test:	88.63	(M_1)	83.24	(M_2)
SD on Form B, final test:	24.36	(σ_1)	21.62	(σ_2)
Gain, $M_1 - M_2$:		5.39		
Standard errors of means, final tests:	2.89		2.57	

Correlation between final scores (experimental and control groups) = .65

The means and σ's of the control and experimental groups in Form A (initial test) are almost identical, showing the original pairing to have been quite satisfactory. The correlation be-

tween the final scores on Form B of the Otis Test is found from the scores of those children who were matched in terms of initial score.*

The difference (D) in the final mean test performance of the experimental and control groups is 88.63 − 83.24 or 5.39. The standard error of this D, σ_D, is found from formula (33) as follows:

$$\sigma_D = \sqrt{(2.89)^2 + (2.57)^2 - 2 \times .65 \times 2.89 \times 2.57} = 2.30$$

The t is 5.39/2.30, or 2.34, and there are 71 degrees of freedom. From Table 29 we find that the incentive group is significantly superior to the control at the $P/2 = .05$ level ($t = 1.67$), and almost significantly superior at the $P/2 = .01$ level ($t = 2.38$). Note that had no account been taken of the correlation between final scores in control and experimental groups, i.e., if formula (29) had been used, the σ_D would have been 3.87. The t would then have been 1.39 instead of 2.34, and the mean gain (*plus* difference) would not have been significant even at the .05 level. Evidently it is very important that we take account of the correlation between final test scores in the experimental and control groups.

When two equivalent groups are small (say 8, 10, or less), a good plan is to compute the differences between final scores made by the paired subjects and follow the method for testing the significance of a mean gain outlined on page 210. The degrees of freedom are one less than the number of pairs.

(*c*) Matched Groups

Investigators often employ the method of *matched groups* when it is not feasible to set up equivalent groups in which subjects have been paired person for person. Groups are matched when they are made alike as regards mean and *SD* in some measure, the matching variable usually being different from the one under study. No attempt is made to pair off individuals and the two groups are not necessarily of the same size,

* Note that the correlation between the final scores of equivalent groups is analogous to the correlation between initial and final scores of the same group. The control group furnishes the "initial" scores.

although a large difference in N is not advisable. In evaluating the final scores of matched groups the procedure is somewhat different from that used in the equivalent groups method.* Let X be the function or test under study, and Y be a variable in terms of which our two groups have been matched as to mean and SD. Then if r_{xy} is the correlation between X and Y in the population from which our matched samples are drawn, the standard error of the difference between means in X is

$$SE_{D_{M_1-M_2}} = \sigma_D = \sqrt{(\sigma^2_{M_{x_1}} + \sigma^2_{M_{x_2}})(1 - r^2_{xy})} \tag{34}$$

(*standard error of the difference between the* M's *of matched groups*)

An example will illustrate the use of this formula.

Example (*7*) The achievement of two groups of first-year high-school boys, the one from an academic, the other from a technical high school, is compared upon a Mechanical Ability Test. The two groups are matched for mean and SD upon a general intelligence test so that the experiment becomes one of comparing the mechanical ability scores of two groups of boys of "equal" general intelligence enrolled in different curricula. Data are as follows:

TABLE 31

	Academic	Technical
No. of boys in each group:	125	137
Means on Intelligence Test (Y):	102.50	102.80
σ's on Intelligence Test (Y):	33.65	28.62
Means on Mechanical Ability Test (X):	51.42	54.38
σ's on Mechanical Ability Test (X):	6.24	7.14

Correlation between the General Intelligence Test and the Mechanical Ability Test for first-year high-school boys is .30.

$$M_{x_1} - M_{x_2} = 54.38 - 51.42 = 2.96$$

By (24a) and (34) $$\sigma_D = \sqrt{\left(\frac{(6.24)^2}{125-1} + \frac{(7.14)^2}{137-1}\right)(1 - .30^2)}$$

$$= .79$$

$$t \text{ or } CR = \frac{2.96}{.79} = 3.75$$

* Lindquist, E. F., "The Significance of a Difference between 'Matched' Groups," *Journal of Educational Psychology*, 22 (1931), 197–204. Wilks, S. S., "The Standard Error of the Means of 'Matched' Samples," *Journal of Educational Psychology*, 22 (1931), 205–208.

Since the degrees of freedom $124 + 136 - 1$ * are quite large, we may take the t values in the bottom line of Table 29 — i.e., assume that the sampling distribution of CR's or t's is normal. Our critical ratio of 3.75 exceeds the .01 level (2.58) and the mean difference is, therefore, highly significant. We may assert with great confidence that boys in the technical high school are definitely better on the Mechanical Ability Test than boys of "equal" verbal intelligence in the academic high school.

The correlation term is introduced in formula (34) because when two groups are matched in one function, their variability (SD) is restricted in those functions correlated with the matching test. For example, height and weight are highly correlated in nine-year-old children. Hence, if a group of nine-year-olds of the same or nearly the same height is selected, the variability in weight of this group will be substantially reduced as compared with nine-year-olds in general. When groups are matched for several variables, e.g., age, intelligence, socio-economic status and the like, and compared with respect to a correlated variable, the correlation coefficient in formula (34) becomes a multiple correlation coefficient (p. 423).

Matched groups and more often equivalent groups have been employed in a variety of psychological and educational studies. Well-known illustrations are found in experiments designed to evaluate the relative merits of two methods of teaching, to determine the effect of drugs, e.g., tobacco or caffeine, upon efficiency, to investigate the transfer effects of special training and many other factors. If the critical ratio (t) in such studies is significant when formula (29) is used, we may have confidence in our result, since the standard error given by formula (29) is *always larger* than the standard error obtained from formula (33) when r is *positive*. If the difference when formula (29) is used is *not* significant, however, it is still possible that it might prove to be so if the experiment were repeated under conditions changed so as to permit the calculation of the correlation between final scores.

* *One* degree of freedom is subtracted for each variable (here one) in terms of which the groups are matched.

2. The Reliability of the Difference between Medians

The reliability of the difference between two medians may be found from the following formula:

$$\sigma_D \text{ or } \sigma_{Mdn_1 - Mdn_2} = \sqrt{\sigma^2_{Mdn_1} + \sigma^2_{Mdn_2}} \tag{35}$$

(*Standard error of the difference between uncorrelated medians*)

3. The Reliability of the Difference between Standard Deviations

(1) The Standard Error of a Difference When σ's Are Uncorrelated

In many studies in psychology and education, the differences in variability which appear between groups are a matter of prime importance. The student of race and sex differences, for example, is often more interested in knowing whether his groups differ significantly in *SD* than in knowing whether they differ in mean score. In like manner, the educational psychologist who is investigating a new method of teaching often wants to know whether his "new" method has produced changes in variability greater than those brought about by the "old" method.

When different groups are studied, or when the tests given to the same group are uncorrelated, the reliability of an obtained difference may be found by the formula

$$\sigma_{D_\sigma} \text{ or } \sigma_{\sigma_1 - \sigma_2} = \sqrt{\sigma^2_{\sigma_1} + \sigma^2_{\sigma_2}} \tag{36}$$

(*standard error of the difference between uncorrelated σ's*)

where σ_{σ_1} is the standard error of the first σ and σ_{σ_2} is the standard error of the second σ (p. 194).

We may apply this formula to the problem of Norwegians and Belgians on page 198. The *SD* of the Norwegians' scores on the combined scale was 2.47; of the Belgians' scores on the same test 2.42. Is this difference in variability significant? Calling the *SD* of the Norwegians' scores σ_1 and the *SD* of the Belgians' scores σ_2, we have, using large sample methods,

$$\sigma_{\sigma_1} = \frac{2.47}{\sqrt{2 \times 611}} = .071 \qquad \text{by (27)}$$

$$\sigma_{\sigma_2} = \frac{2.42}{\sqrt{2 \times 129}} = .151 \qquad \text{by (27)}$$

$$\sigma_{D\sigma} = \sqrt{(.071)^2 + (.151)^2} = .167 \text{ or } \mathbf{.17} \qquad \text{by (36)}$$

The obtained difference in the σ's is 2.47 — 2.42 or .05. Dividing this difference by .17 $t = .05/.17$ or 1.30. On the null hypothesis (Table 17), differences larger than $\pm\ 1.30\sigma_{D\sigma}$ can be expected to occur about eight times in ten trials from sampling errors alone. The given difference is clearly not significant, therefore, and the null hypothesis is retained.

(2) The Standard Error of a Difference When σ's Are Correlated

When we compare the *SD*'s of the same group upon two occasions or the *SD*'s of equivalent groups on a final test, we must take into account the correlation between the *SD*'s of the groups being compared. The formula for testing the significance of an obtained difference in variability when *SD*'s are correlated is

$$\sigma_{D\sigma} = \sqrt{\sigma^2_{\sigma_1} + \sigma^2_{\sigma_2} - 2r^2_{12}\sigma_{\sigma_1}\sigma_{\sigma_2}} \qquad (37)$$

(*standard error of the difference between correlated* σ's)

where σ_{σ_1} and σ_{σ_2} are the standard errors of the two *SD*'s and r^2_{12} is the *square* of the coefficient of correlation between scores in final and initial tests of the same group or between final scores of equivalent groups.*

Formula (37) may be applied to the problems on pages 209 and 211. In the first problem (p. 209) the *SD* of the sixty-five sixth-grade children is 6.0 on the initial test and 5.0 on the final test. Is there a significant difference in variability in reading

* The correlation between the *SD*'s of samples drawn from a given population equals the square of the coefficient of correlation between the test scores, the *SD*'s of which are being compared. See Kelley, T. L., *Statistical Method* (1923), p. 178.

after a year's schooling? If we call $\sigma_1 = 6.0$, and $\sigma_2 = 5.0$, we have

$$\sigma_{\sigma_1} = \frac{6.0}{\sqrt{2 \times 64}} = .53 \qquad \text{by (27a)}$$

$$\sigma_{\sigma_2} = \frac{5.0}{\sqrt{2 \times 64}} = .44 \qquad \text{by (27a)}$$

The coefficient of correlation between initial and final scores is .60, so that $r^2 = .36$. Substituting for r^2 and the σ_σ's in formula (37), we have

$$\sigma_{D\sigma} = \sqrt{(.53)^2 + (.44)^2 - 2 \times .36 \times .53 \times .44}$$
$$= .55$$

The difference between the σ's divided by .55 $\left(\frac{6-5}{.55}\right) = 1.80$. The t for 64 degrees of freedom is 2.00 at $P = .05$. The computed t of 1.80 does not quite reach this point. Hence there is no reason for believing that a real difference in variability exists as between these two groups.

In the equivalent groups problem (p. 211) the *SD* of the experimental group on the final test was 24.36, and the *SD* of the control group on the final test was 21.62. The difference between these *SD*'s is 2.74, and the number of children in each group is seventy-two. Did the incentive (praise) produce significantly greater variability in the experimental group as compared with the control? Putting $\sigma_1 = 24.36$ and $\sigma_2 = 21.62$, we have

$$\sigma_{\sigma_1} = \frac{24.36}{\sqrt{2(72-1)}} = 2.04 \qquad \text{by (27a)}$$

$$\sigma_{\sigma_2} = \frac{21.62}{\sqrt{2(72-1)}} = 1.81 \qquad \text{by (27a)}$$

The coefficient of correlation between final test scores in the experimental and control groups is .65, and r^2_{12} is .42. Substituting for r^2 and the standard errors in formula (37) we have

$$\sigma_{D\sigma} = \sqrt{(2.04)^2 + (1.81)^2 - 2 \times .42 \times 2.04 \times 1.81}$$
$$= 2.08$$

If we divide 2.74 by 2.08 our critical ratio or t is 1.32. For 71 degrees of freedom this t (Table 29) is not significant at the $P = .05$ level (2.00) nor in the positive direction at the $P/2 = .05$ level (1.67). There is no evidence, then, that the incentive increased the variability of response.

V. The Reliability of Certain Other Measures

This section will consider the standard errors of certain statistics which are used fairly often in experimental work. The reliability of r, the coefficient of correlation, will be treated in Chapter IX, page 297. For the standard errors of many other important measures the student should go to the more advanced references in the literature. *The Handbook of Statistical Nomographs, Tables, and Formulas*, by Dunlap and Kurtz, contains many formulas which are often needed in research investigations.

1. The Standard Error of a Percentage and the Standard Error of the Difference between Two Percentages

It is often possible to find the percentage of a given group which exhibits a certain attribute or possesses certain interests or attitudes, or other fairly general characteristics, when it is difficult if not impossible to measure these attributes directly. Given the percentage occurrence of an attribute, the question of how much confidence we can put in our figure often arises. How reliable an index is it of the incidence of the phenomenon in which we are interested? The standard error of a percentage is given by the formula

$$\sigma_{\%} = 100\sqrt{\frac{pq}{N}} = 100\sqrt{\frac{p(1-p)}{N}} \qquad (38)$$

(*standard error of a percentage*)

in which p = the proportion of times the given event occurs; $q = 1 - p$; and N = the number of cases.

We may illustrate this formula with a problem:

Example (1) In a study of cheating, a group of 613 elementary school children were classified as to the occupations

of their fathers. It was found that 348 children had fathers who were professional men, business men, merchants, etc. Of these 348 children of "good" social status, 144 or 41.4% were found to have cheated on various tests given in school. Assuming our sample to be representative of children from the given social level, how much confidence may be placed in the stability of this percent? How much fluctuation in percent cheating might be expected if we investigated a number of groups of children whose fathers fall into the same occupational classification? *

Applying formula (38), we get

$$\sigma_{\%} = 100\sqrt{\frac{.414 \times .586}{348}} = 2.7\%$$

This standard error is interpreted as is σ_M for large samples; that is, we assume the sampling distribution of CR's to be normal. On the evidence, therefore, the probability is .95 that the percentage of children cheating really lies between 46.7% and 36.1% ($41.4 \pm 1.96 \times 2.7$). Only five times in 100 trials would we expect a percentage to occur outside of these limits.

We often want to know whether there is a significant difference between the percentages of two groups who exhibit a certain form of behavior. When our two groups constitute samplings from what seem to be different populations, or when percentages are uncorrelated,† we may determine the significance of the difference between the percentages in the two groups by the formula:

$$\sigma_{D\%} = \sqrt{\sigma^2_{\%_1} + \sigma^2_{\%_2}} \tag{39}$$

or

$$\sigma_{D\%} = 100\sqrt{\frac{p_1q_1}{N_1} + \frac{p_2q_2}{N_2}}$$

(standard error of the difference between two uncorrelated percentages)

* Hartshorne, H., and May, M. A., *Studies in Deceit* (1928), Book II, 161.

† If certain members of Group I are more likely to cheat, when certain members of Group II cheat, percentages cheating in the two groups will be correlated.

We may illustrate the use of this formula by reference to Example (1) given above. It was stated that 41.4% of the 348 children, classified as of "good" social status, cheated on the tests given. In the same study, 50.2% of 265 children whose fathers were classified as skilled and unskilled laborers, i.e., were of relatively "poor" social status, cheated on the same tests of deception. Is there a "real" difference in "deceptive behavior" between these two groups? The $\sigma_{\%}$ for the percentage .502 in the second group is

$$\sigma_{\%} = 100\sqrt{\frac{.502 \times .498}{265}} = 3.1\% \quad \text{by} \qquad (38)$$

Calling 2.7 $\sigma_{\%_1}$, and 3.1 $\sigma_{\%_2}$, and substituting in formula (39), we have

$$\sigma_{D_{\%}} = \sqrt{2.7^2 + 3.1^2} = 4.1\%$$

The difference between the percentages of those who cheated in the two groups is 50.2 − 41.4 or 8.8. Dividing 8.8 by 4.1, we obtain a *CR* of 2.15. Assuming the distribution of *CR*'s to be normal (samples are large), we find from the bottom line of Table 29 that a t of 2.15 is significant at the .05 level (1.96), but not at the .01 level (2.58).

2. The Standard Errors of Measures of Skewness and of Kurtosis

(1) Skewness

In Chapter V, page 121, a formula for estimating the skewness of a frequency distribution in terms of its median and certain percentiles was given as follows:

$$Sk = \frac{(P_{90} + P_{10})}{2} - P_{50} \qquad (19)$$

According to this formula, the skewness of the 50 Army Alpha scores (the distribution is given in Table 1, p. 6), is − 2.50.

The significance of this measure of skewness may be determined by means of the formula

$$\sigma_{sk} = \frac{.5185D}{\sqrt{N}} \tag{40}$$

(*standard error of the measure of skewness given in formula (19)**)

in which $D = (P_{90} - P_{10})$.

In the frequency distribution of 50 Army Alpha scores, P_{90} is 187, P_{10} is 152, and $D = 35$. From formula (40), therefore,

$$\sigma_{sk} = \frac{.5185 \times 35}{\sqrt{50}} = 2.57$$

and dividing -2.50 (Sk) by 2.57 (σ_{sk}), we get a t of .97 (the sign of Sk indicates the *direction* of skewness). Assuming the distribution of t's to be normal, it is clear from Table 29 that this t falls far short of the .05 level. We may feel quite sure, then, that the distribution is *not* significantly skewed.

The skewness of the distribution of 200 cancellation scores (p. 14), is, by formula (19), .03; $P_{90} = 128.5$, $P_{10} = 110.4$, and $D = 18.1$. The standard error of Sk is

$$\sigma_{sk} = \frac{.5185 \times 18.1}{\sqrt{200}} = .66$$

Dividing .03 (Sk) by .66 (σ_{sk}), we get a t of .046; and from Table 29 find that the skewness is far from being significant. In fact this distribution is almost perfectly symmetrical (Fig. 5, p. 20, verifies this result).

(2) Kurtosis

On page 122 the following formula was given for measuring the kurtosis of a distribution in terms of Q and certain percentiles:

$$Ku = \frac{Q}{(P_{90} - P_{10})} \tag{20}$$

* Kelley, T. L., *Statistical Method* (1923), p. 77. The formula, as given in this reference, is in error; see Dunlap, J. W., and Kurtz, A. K., *Handbook of Statistical Nomographs, Tables and Formulas* (1932), p. 112.

The kurtosis of the frequency distribution of 50 Army Alpha scores by formula (20) is given on page 122 as .237. This value deviates $-.026$ from the Ku of the normal distribution which is .263 (to three decimals). The *direction* of the deviation indicates that the distribution is leptokurtic.

We may estimate the significance of our deviation of $-.026$ from "normal" kurtosis by calculating σ_{Ku}, using the following formula:

$$\sigma_{Ku} = \frac{.27779}{\sqrt{N}} \tag{41}$$

(standard error of the measure of Ku *given by formula (20))*

in which N is the size of the sample.

For the fifty Army Alpha scores, $\sigma_{Ku} = \dfrac{.27779}{\sqrt{50}} = .039$, and t or $Ku_d/\sigma_{Ku} = .026/.039$ or .67. Assuming a normal sampling distribution for t, .67 is well below the .05 level (Table 29) and the deviation ("peakedness") of this frequency distribution from the normal form is not significant.

The kurtosis of the 200 cancellation scores (p. 122) is by formula (20) .223, a value which deviates $-.040$ from .263, the Ku of the normal distribution. The *direction* of the deviation indicates leptokurtosis.

To determine the significance of this deviation from normality, calculate σ_{ku} which equals .020. Ku_d/σ_{Ku} equals .040/.020 or 2.00, and from Table 29 we find that the leptokurtosis of the distribution is significant at the .05 level ($P/2 = 1.65$) but not at the .01 level. The narrow dispersion of this distribution ($Q = 4.04$), leading to a concentration of cases in the middle range, probably accounts for its strong tendency to be more "peaked" than the normal distribution (see p. 122).

VI. Sampling and the Use of Reliability Formulas

All of the reliability formulas given in this chapter depend upon N, the number of cases in the sample, and most of them

involve some measure of variability (usually σ) calculated from the data. It is unfortunate, perhaps, that given these statistics there is nothing in the statement of a reliability formula itself which might deter the uncritical worker from applying it to any set of test scores. General and indiscriminate calculation of standard errors, however, will lead to erroneous conclusions and false interpretations. For this reason, it is important that the research worker in experimental psychology or in education have clearly in mind (1) the conditions under which reliability formulas are — and are not — applicable; and that he know (2) what his formulas may reasonably be expected to do.* Some of the limitations to reliability formulas have been pointed out in this chapter. These statements will now be amplified and certain cautions to be observed in the use of reliability formulas indicated.

1. Reliability Formulas Assume Random Samples

Reliability formulas apply strictly to random samples only: when other sampling methods have been employed, special techniques must be used in determining significance levels.† The criterion of randomness in a sample is met when every person in the population from which the sample has been drawn has had an equal chance of being chosen. A random sample is truly representative of its population, since cases are chosen without bias as to able, mediocre, and poor individuals. It may seem paradoxical, but one must often take great pains to "select" his sample randomly. To be representative of ten-year-old boys within a given city, for example, a group must not be drawn exclusively from a poor neighborhood, from expensive private schools, or from any larger group in which special factors are known to play an important rôle.

Mental traits which have been carefully measured in large samples have usually proved to be normally or approximately

* Walker, Helen M., *Elementary Statistical Method* (1943), Chapter 15, pp. 263–271.

† McNemar, Q., *Sampling in Psychological Research*, Psychological Bulletin, 37 (1930), 331–365.

normally distributed. We may make the reasonable assumption, therefore, that many of the traits in which we are interested follow the normal distribution in the general population. Random samples drawn from a normally distributed population will also be normally distributed, so that normality becomes one criterion of adequacy in a sample. The range covered by samples of different sizes (all drawn from a normal population) will be approximately as follows:

$N = 10$	Range $\pm$ 2.0σ
$N = 50$	Range $\pm$ 2.5σ
$N = 200$	Range $\pm$ 3.0σ
$N = 1000$	Range $\pm$ 3.5σ

A range of $\pm$ 3.5σ from the mean includes, in a normally distributed group, 9995 cases in 10,000 (Table 17). The same range includes, of course, 99.95% of the cases in a sample of 100. In the sample of 10,000, five cases fall outside of this range; in a sample of 100, no cases lie outside of the given range. The more extreme the deviation, the less the probability of its occurrence; and in small samples, wide deviations from the mean rarely appear if the sample is truly representative of a normally distributed group. When working with small samples, therefore, deviations far removed from the mean should often be discarded much as a laboratory worker throws out measures of reaction time which are obviously premature or delayed.

One of the simplest tests of the adequacy — the representativeness — of a sample consists in drawing from the population another group of approximately the same size as the sample with which we are working. If the means and sigmas computed from these two independently drawn groups are of almost the same size, we may feel reasonably sure that *both* samples are representative of the population. If the correspondence is not close, we may try the expedient of adding new cases to our samples until they yield means and σ's which are increasingly similar or increasingly dissimilar. In the latter event neither sample is likely to be adequate. More information may be

secured with respect to the reliability of a mean or σ by repeated sampling, or by a careful study of several samples, than can be obtained from an uncritical and blanket use of reliability formulas.

2. Reliability Formulas Assume a "Sufficiently Large" Sample

The value of a standard error is conditioned, in part at least, upon our having a sufficiently large sample. A small sample may be satisfactory in intensive laboratory studies in which many measurements are taken on each subject. But if N is less than about 25, there is usually little reason for assuming such a small sample to be descriptive of a given population. As we have seen (p. 183) standard errors vary inversely as the size of the sample; hence, the larger the sample in general the smaller the error. A fairly simple and practical method of deciding when a sample is "sufficiently large" is to increase N until the addition of extra cases drawn at random fails to produce an appreciable change in the mean or σ. When this point is reached, the sample is probably large enough to be taken as descriptive of its population. But the corollary must be recognized that mere numbers do not in themselves guarantee a representative sample.*

3. Reliability Formulas Measure Fluctuations Arising from Sampling and from Errors of Measurement

Standard errors of means, σ's, etc., measure both (1) sampling errors, and (2) errors of measurement, i.e., variable errors in the test scores themselves (p. 394). We have already considered the question of the sampling error of the mean on page 184. If a sample were perfectly representative of its population, its mean and σ would equal the mean and σ of the population. Except by chance, however, neither a given sample nor another similarly selected and approximately of the same size will describe the

* See *The New Science of Public Opinion Measurement* (American Institute of Public Opinion, Princeton, N. J.).

entire population perfectly. Moreover it is unlikely that means calculated from successive samples will equal each other. Uncertainty as to the reliability of a calculated measure grows out of the fact that we must necessarily work with samples instead of with the whole population. Variations from sample to sample — the so-called "errors of sampling" — are not to be thought of as mistakes, failures and the like, but rather as fluctuations which arise from the fact that no two samples are ever exactly alike. If samples are random and sufficiently large, and if there is no constant error, calculated means will tend to vary around the true mean of the population within a comparatively small range. This range is given by the standard error. The accuracy limits of a mean (p. 187) should be calculated from the t-distribution (Table 29) when N is small, and from the normal probability distribution when N is large.

If the standard error of a mean is large, it does not follow necessarily that the mean is affected by a large sampling error. Much of the error may be due to errors of measurement. On the other hand, when errors of measurement are known to be negligible, a small standard error does indicate that the reliability of a calculated measure is high insofar as sampling fluctuations are concerned. In other words, when the standard error is small a mean or σ is a good estimate of the population mean or σ.

4. Reliability Formulas Do Not Measure the Effects of Constant Errors Nor the Failure to Get a Random Sample

Errors which arise from inadequate sampling are neither detected nor measured directly by reliability formulas. For example, the mean score on an intelligence test made by 500 male college students between the ages of eighteen and twenty-five will not be representative of the achievement of the male population within this age range. College students constitute a highly selected group; and in consequence, other samples of 500 young men, aged eighteen to twenty-five, and drawn at random from the male population will return very different

means and sigmas from those obtained with the college group. These differences in mean and σ cannot be attributed to sampling errors, since samples were not drawn at random from the same population. If our population were restricted to college men, our original sample of 500 might, of course, be entirely adequate.

Reliability formulas are affected by, but do not reveal, constant errors. Constant errors work in only one direction, are always plus or always minus. Constant errors arise from many sources — familiarity with test material, fatigue, faulty technique in giving and scoring tests (over- and under-timing are examples), in fact, from a consistent bias of almost any sort. Standard errors calculated for measures subject to such influence when not definitely misleading are at best of doubtful values. The careful study of successive samples, rechecks when practicable, care in controlling conditions, and the use of objective checks will eliminate many prolific and troublesome sources of constant error. The research worker should always bear in mind that even the most refined statistical technique cannot make bad data yield valid results.

PROBLEMS

1. Given: $M = 26.40$; $\sigma = 3.20$; $N = 100$.
   (*a*) Determine the accuracy limits of this M at the .05 level; at the .01 level.
   (*b*) Determine the accuracy limits of σ at the .05 level and .01 level.
2. Given: $Mdn = 72.40$; $Q = 12.84$; $N = 81$.
   (*a*) Determine the accuracy limits of Mdn at the .05 level; at the .01 level.
   (*b*) Determine the accuracy limits of Q at the .05 level.
3. The mean of a large sample is K and σ_K is 2.50. What are the chances that the sample mean misses the true mean by more than (*a*) $\pm$ 1.00; (b) $\pm$ 3.00; (c) $\pm$ 10.00?
4. The following five measures of perception span for unrelated words are obtained from one observer:

   5 6 4 7 5

(a) Determine .05 and .01 accuracy limits for the mean (page 201).
(b) Determine .05 accuracy limits for the SD.
(c) Compare the .05 accuracy limits for the mean when calculated by large and by small sample methods.

5. The difference between two means ($M_1 - M_2$) is 3.60, the $\sigma_D =$ 3.00 and the samples are large.

   (a) Is the obtained difference significant at the .05 level?
   (b) What percent is the obtained difference of the difference necessary for significance at the .01 level?

6. A personality inventory is administered in a private school to eight boys whose conduct records are exemplary, and to five boys whose records are very poor. Data are given below.

Group 1:	110	112	95	105	111	97	112	102
" 2:	115	112	109	112	117			

   Is the difference between group means significant at the .05 level? at the .01 level?

7. In the first trial of a practice period, twenty-five twelve-year-olds have a mean score of 80.00 and a σ of 8.00 upon a digit-symbol learning test. On the tenth trial, the mean is 88.00 and the σ is 10.00. The r between scores on the first and tenth trials is .40.

   (a) Is the gain in score significant at the .05 level? at the .01 level?
   (b) Is the increase in variability significant at the .05 level? at the .01 level?

8. Two groups of high-school pupils are matched for initial ability in a biology test. Group 1 is taught by the lecture method, and Group 2 by the lecture-demonstration method. Data are as follows:

	Group 1 (control)	Group 2 (experimental)
N	60	60
Mean initial score on the biology test	42.30	42.50
σ of initial scores on the biology test	5.36	5.38
Mean final score on the biology test	54.54	56.74
σ of final scores on the biology test	6.34	7.25
r (between final scores on the biology test) = .50		

(a) Is the difference between the final scores made by Groups 1 and 2 upon the biology test significant at the .05 level? at the .01 level?

(b) Is the difference in the variability of the final scores made by Groups 1 and 2 significant at the .05 level?

9. Two groups of high-school students are matched for M and σ upon a group intelligence test. There are fifty-eight subjects in Group A and seventy-two in Group B. The records of these two groups upon a battery of "learning" tests are as follows:

	Group A	Group B
M	48.52	53.61
σ	10.60	15.35
N	58	72

The correlation of the group intelligence test and the learning battery in the entire group from which A and B were drawn is .50. Is the difference between Groups A and B significant at the .05 level? at the .01 level?

10. Calculate measures of skewness and kurtosis for each of the four distributions in Chapter II, problem 1, page 46. Compute standard errors of Sk and Ku by the formulas given on pages 221 and 222. Determine whether any of these distributions departs significantly from the normal form.

11. In a city high school of 5000 pupils, 52.3% are girls; and in a second high school of 3000 pupils, 47.7% are girls. Is there a significant difference between the percentages of girls enrolled in the two high schools?

12. In an institution, eighty delinquent and eighty non-delinquent boys of the same age, same I.Q., and roughly the same social status furnish the following data:

(a) 40% of the delinquent, and 20% of the non-delinquent come from "poor" homes.

(b) 74% of the delinquent and 44% of the non-delinquent score above the "normal" median on a neurotic inventory.

(c) 65% of the delinquent and 50% of the non-delinquent cheat on a certain test.

Are any of these differences significant?

13. In a random sample of 100 cases each from four groups, A, B, C and D, the following results were obtained:

	A	B	C	D
Mean	101.00	104.00	93.00	86.00
σ	10.00	11.00	9.60	8.50

What are the chances that, in general, the mean of

(*a*) the B's is higher than the mean of the A's.
(*b*) the A's is higher than the mean of the C's.
(*c*) the C's is higher than the mean of the D's.

What are the chances that

(*a*) a B will be better than the mean A.
(*b*) a B will be better than the mean C.
(*c*) a B will be better than the mean D.

Answers

1. (*a*) 25.77 and 27.03; 25.57 and 27.23
   (*b*) 2.75 and 3.65; 2.61 and 3.79
2. (*a*) 67.21 and 77.59; 65.60 and 79.20
   (*b*) 9.59 and 16.09
3. 69 in 100; 23 in 100; less than 1 in 100
4. (*a*) 3.98 and 6.82; 3.05 and 7.75
   (*b*) .14 and 2.14
   (*c*) 4.50 and 6.30; 3.98 and 6.82
5. (*a*) No. $CR = 1.20$
   (*b*) 46.5%
6. $t = 2.3$; significant at .05 but not at .01 level
7. (*a*) (D/σ_D) or $t = 3.92$; significant at .05 and at .01 levels
   (*b*) $(D/\sigma_{D\sigma})$ or $t = 1.18$; not significant at .05 level
8. (*a*) $t = 2.47$; significant at .05 but not at .01 level
   (*b*) $t = 1.18$; not significant at .05 level
9. $t = 2.57$; significant at .05 and at .01 levels
10.

Distribution	Sk/σ_{sk}	Ku_d/σ_{ku}	
1	−.23	.55	Deviation from normality not significant
2	.51	−.38	" " " " "
3	.33	.93	" " " " "
4	.13	.68	" " " " "

11. $D/\sigma_{D_{\%}} = 4.0$; significant at .01 level

12. (*a*) $D/\sigma_{D_{\%}} = 2.83$; significant at .01 level
    (*b*) $D/\sigma_{D_{\%}} = 4.05$; significant at .01 level
    (*c*) $D/\sigma_{D_{\%}} = 1.94$; almost significant at .05, not at .01, level

13. (*a*) 98 in 100
    (*b*) more than 99 in 100
    (*c*) more than 99 in 100

    (*a*) 61 in 100
    (*b*) 84 in 100
    (*c*) 95 in 100

CHAPTER VIII

TESTING EXPERIMENTAL HYPOTHESES

A PSYCHOLOGICAL experiment is designed to answer some question which the investigator has in mind. The investigator's hypothesis may be in the nature of a general proposition or it may be a specific query. A specific hypothesis is, ordinarily, to be preferred to a general one, as the more definite and exact the thesis the greater the likelihood of a conclusive answer. In the preceding chapter we were concerned with testing hypotheses concerning differences of various sorts: differences between means, σ's, percentages, and the like. The significance of obtained differences was tested by calculating a critical ratio which was evaluated in terms of the normal distribution (p. 115) or the t-distribution (p. 190). In the present chapter we shall consider somewhat more carefully the nature of hypotheses and shall present certain useful ways of answering the questions raised by an experiment.

I. THE NULL HYPOTHESIS

1. Meaning of the Null Hypothesis

We have already had occasion to employ the null hypothesis in Chapter VII, where the significance of the differences between two groups was to be tested. The null hypothesis, it will be remembered, asserts that *no* true difference exists as between our two samples; that, in fact, these samples were randomly drawn from the same population, and differ only by accidents of sampling. A null hypothesis, therefore, constitutes a *challenge;* and the function of an experiment is to give the facts a chance to meet (or fail to meet) this challenge. To illustrate, suppose it has been claimed that ten-year-old girls read better than ten-year-old boys. This hypothesis is indefinite as it stands, and hence is not testable, as we do not know how much better than boys the girls must read before they can be said to

"read better." If we assert that girls read *no* better than boys or — to say the same thing — that such differences as are found in reading ability as between groups of ten-year-old boys and girls can be attributed to accidents of sampling, this (null) hypothesis *is* exact and can be tested by the usual sampling formulas. Suppose that groups of ten-year-old boys and girls are drawn at random from the school population, and that on a standard reading examination the mean score of girls is significantly higher than the mean score of boys. If this happens the null hypothesis is disproved and must be rejected. In discarding the null hypothesis what we are really saying is that the difference in reading achievement as between boys and girls cannot be fully explained by sampling fluctuations.

It is important to realize that the rejection of a null hypothesis does not *force* the acceptance of a contrary view.* A significant difference in reading ability as between ten-year-old boys and girls, for instance, does not *prove* girls to be better readers, it simply means that the two groups do actually differ. In subsequent comparisons of boys and girls, if *all* experimental variables likely to influence the reading score are controlled and the difference still remains, we may then be willing to assert the existence of a true sex difference in reading ability. But the acceptance of a positive hypothesis — it should be noted — is usually the end result of a series of experiments. Furthermore, it is a logical and not a statistical conclusion.

The extra-sensory perception (ESP) experiments† offer a good illustration of the meaning of a null hypothesis. In a typical experiment in ESP a pack of twenty-five cards is used. There are five different symbols on these cards, each symbol appearing on five cards. In guessing through a pack of cards, the probability of chance success with each card is 1/5 (on the average). And the number of correct "calls" in a pack of twenty-five cards should be five. If a subject calls the cards

* Morgan, J. J. B., "Credence Given to One Hypothesis because of the Overthrow of Its Rivals," *American Journal of Psychology*, 58 (1945), 54–64.

† Rhine, J. B., et al., *Extra-Sensory Perception after Sixty Years* (New York: Henry Holt and Co., 1940).

correctly considerably in excess of chance expectation (i.e., in excess of five), the null hypothesis is rejected. But rejection of the null hypothesis does not force immediate acceptance of ESP as the *cause* of extra-chance results. Before this conclusion can be reached we must demonstrate in a series of experiments that extra-chance results are obtained when we have eliminated *all* likely causes such as runs of cards, cues, poor shuffling and recording, and the like. If under rigid controls results in excess of chance are consistently achieved, we may reject the null hypothesis and accept ESP. But the acceptance of ESP, as of any positive hypothesis, is necessarily tentative and is contingent upon further work.

Ordinarily, the null hypothesis is more useful than other hypotheses because it is *exact*. Hypotheses which assert that some group is "better" or "more accurate" or "more skilled" than another are inexact and cannot be tested, as we cannot quantify our expected finding. Hypotheses other than the null hypothesis can, to be sure, be made exact: we may, for example, assert that a group which has received special training will be *five* points on the average better than an untrained (control) group. It is difficult, however, to set up such precise expectations in most experiments; and for this reason it is advisable to adopt the null hypothesis in preference to others if this can be done.

2. Testing the Null Hypothesis against the Direct Determination of Probable Outcomes

The null hypothesis can often be efficiently tested by comparing experimentally observed results with those to be expected from probability theory. Several examples will illustrate the methods to be employed.

Example (1) Two tones, differing slightly in pitch, are to be compared in an experiment. The tones are presented in succession, the subject being instructed to report the second as higher or lower than the first. Presentation is in random order. In ten trials a subject is right in his judgment seven times. Is this result significant, i.e., better than chance?

Since the subject is either right or wrong in his judgment, and since judgments are separate and independent, we may test our result against the binomial expansion (p. 104). Ten judgments may be taken as analogous to ten coins; a right judgment corresponds to a head, say, a wrong judgment to a tail. The odds are even that any given judgment will be right; hence in ten trials (since $p = 1/2$) our subject should in general be right five times by chance alone. The question, then, is whether seven "rights" are significantly greater than the expected five. From page 108 we find that upon expanding $(p + q)^{10}$ the probability of ten right judgments is 1/1024; of nine right and one wrong, 10/1024; of eight right and two wrong, 45/1024; and of seven right and three wrong, 120/1024. Adding these fractions we get 176/1024, or .172 as the probability of seven or *more* right judgments by chance alone. The probability of *just* seven rights is 120/1024 or approximately .12. Neither of these results is significant at the .05 level of confidence (p. 201) and accordingly the null hypothesis must be retained. On the evidence there is no reason to believe that our subject's judgments are really better than chance expectation.

Note that to get ten right is highly significant (the probability is approximately .001); to get nine *or* ten right is also significant (the probability is 1/1024 + 10/1024 or approximately .01). To get eight or *more* right is almost significant at the .05 level (the probability is .055); but any number right less than eight fails to reach our standard. The situation described in Example (1) occurs in a number of experiments — whenever, for example, objects, weights, lights, test items, or other stimuli are to be compared, the odds being 50:50 that a given judgment is correct.

Example (2) Ten photos, five of feeble-minded and five of normal children (of the same age and sex), are presented to a subject who claims he can identify the feeble-minded from their photographs. The subject is instructed to designate which five photographs are those of feeble-minded children. How many photos must our subject identify correctly before the null hypothesis is disproved?

Since there are five feeble-minded and five normal photos, the subject has a 50:50 chance of success with each photo and the method of Example (1) could be used. A better test,* however, is to determine the probability that a particular set of five photos (namely, the *right* five) will be selected from all possible sets of five which may be drawn from the ten given photos. To find how many combinations of five photos can be drawn from a set of ten, we may use conveniently the formula for the combination of ten things taken five at a time. This formula† is written $C^{10}_5 = \frac{10!}{5!\,5!} = 252$. The symbol C^{10}_5 is read "the combinations of ten things taken five at a time"; $10!$ (read "10 factorial") is $10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$; and $5!$ is $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$.

It is possible, therefore, to draw 252 combinations of five from a set of ten, and accordingly there is one chance in 252 that a judge will select the five correct photos out of all possible sets of five. If he does select the right five, this result is obviously significant (the probability is approximately .004) and the null hypothesis must be rejected. Suppose that our judge's set of five photos contains four feeble-minded and one normal picture; cr three feeble-minded and two normal pictures. Is either of these results significant? The probability of four right selections and one wrong selection by chance is $\frac{C^5_4 \times C^5_1}{C^{10}_5}$, i.e., the product of the number of ways four rights can be selected from the five feeble-minded pictures times the number of ways one wrong can be selected from the five normal pictures divided by the total number of combinations of five. Calculation shows this result to be 25/252 or 1/10 (approximately) and hence *not* significant at the .05 level. The probability of getting three right and two wrong is given by $\frac{C^5_3 \times C^5_2}{C^{10}_5}$; namely, the product

* Fisher, R. A., *The Design of Experiments* (1935) Chapter 2, pp. 26–29 especially.

† The general formula for the combinations of n things taken r at a time is $C^n_r = \frac{n!}{r!\,(n-r)!}$

of the number of ways three pictures can be selected from five (the five feeble-minded pictures) times the number of ways two pictures can be selected from the five normal pictures divided by the total number of combinations of five. This result is 100/252 or slightly greater than 1/3, and is clearly not significant.

Our subject disproves the null hypothesis, then, *only* when *all* five feeble-minded pictures are correctly chosen. The probabilities of various combinations of right and wrong choices are given below — they should be verified by the reader:

Probability of	all	5R =	1/252
"	"	4R =	25/252
"	"	3R =	100/252
"	"	2R =	100/252
"	"	1R =	25/252
"	"	0R =	1/252

It may be noted that by increasing the number of pictures of feeble-minded and normal from ten to twenty, say, the *sensitiveness* of the experiment can be considerably enhanced. With twenty pictures it is not necessary to get all ten feeble-minded photos right in order to achieve a significant result. In fact, eight right is significant at the .01 level as shown below.

$$C^{20}_{10} = \frac{20!}{10!\,10!} = 184{,}756$$

Combinations		Frequency	Prob. ratio (freq. ÷ 184,756)
10R	0W	1	.000005
9R	1W	100	.0005
8R	2W	2025	.011
7R	3W	14400	.078
6R	4W	44100	.238
5R	5W	63504	.343
4R	6W	44100	.238
3R	7W	14400	.078
2R	8W	2025	.011
1R	9W	100	.0005
0R	10W	1	.000005
		184,756	

3. Testing the Null Hypothesis against Probabilities Calculated from the Normal Curve

When the number of observations or the number of trials is large, direct calculation of probabilities by expanding the binomial $(p + q)^n$ becomes highly laborious. Since $(p + q)^n$ yields a distribution (p. 110) which is essentially normal when n is large, in many experiments the normal curve may be usefully employed to provide expected results under the null hypothesis. An example will make the method clear.

> *Example (3)* In answering a test of 100 true-false items, a subject gets sixty right. Is it likely that the subject merely guessed?

As there are only two possible answers to each item, one of which is right and the other wrong, the probability of a correct answer to any item is 1/2, and our subject should by chance answer 1/2 of 100 or 50 items correctly. Letting p equal the probability of a right answer, and q the probability of a wrong answer, we could, by expanding the binomial $(p + q)^{100}$, calculate the probability of various combinations of rights and wrongs on the null hypothesis. When the exponent of the binomial (here, number of items) is as large as 100, however, the resulting distribution is very close to the normal probability curve (p. 110) and may be so treated with little error.

Figure 46 illustrates the solution of this problem. The mean of the curve is set at 50. The *SD* of the probability distribution found by expanding $(p + q)^n$ is $\sigma = \sqrt{npq}$; hence for $(p + q)^{100}$, $\sigma = \sqrt{100 \times 1/2 \times 1/2}$ or 5. A score of 60 covers the interval on the baseline from 59.5 up to 60.5. The lower limit of 60 is 1.9σ removed from the mean $\left(\frac{59.5 - 50}{5} = 1.9\sigma\right)$; and from Table 17 we find that 2.87% of the area of a normal curve lies *above* 1.9σ. There are only three chances in 100 that a score of 60 (or more) would be made if the null hypothesis were true. A score of 60, therefore, is significant at the .05 level. We may reject the null hypothesis with some confidence

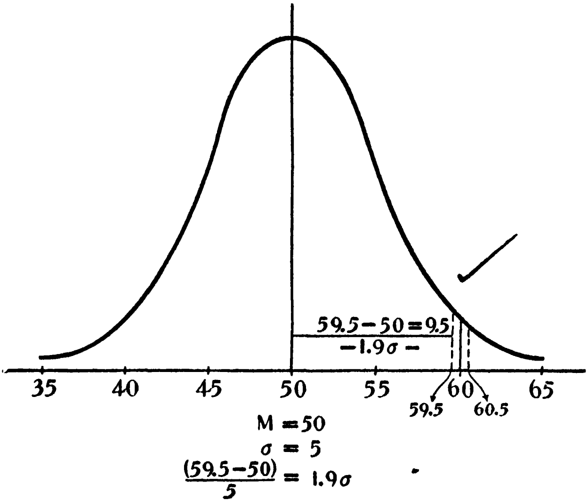


FIG. 46.

and conclude that our subject could not have been simply guessing.

Note that the problem above could have been solved equally well in terms of percentages. We should expect our subject to get 50% of the items right by guessing. The *SD* of this percentage is $100\sqrt{\frac{pq}{n}}$ or $100\sqrt{\frac{.50 \times .50}{100}}$ or 5%. A score of 60% (lower limit 59.5%) is 9.5% or 1.9σ distant from the middle of the curve. We interpret this result in exactly the same way as that above.

> *Example* (*4*) A multiple-choice test of sixty items provides four possible responses to each item. How many items should a subject answer correctly before we may feel sure that he knows something about the test material?

Since there are four responses to each item, only one of which is correct, the probability of a right answer by guessing is 1/4, of a wrong answer 3/4. The final score to be expected if a subject knows nothing whatever about the test and simply guesses

is $1/4 \times 60$ or 15. Our task, therefore, is to determine how much better than 15 a subject must score in order to demonstrate real knowledge of the material.

This problem could be solved by the methods of Example (1). By expanding the binomial $(p + q)^n$ in which $p = 1/4$, $q = 3/4$, and $n = 60$, we can determine the probability of the occurrence of any score from 0 to 60. The direct determination of probabilities from the binomial expansion is straightforward and exact but the calculation is rather tedious. A satisfactory approximation to the answer we want may be obtained by using the normal probability distribution to determine probabilities, as in Example (3). The mean of our "chance" distribution is 1/4 of 60 or 15; and the $\sigma = \sqrt{npq} = \sqrt{60 \times 1/4 \times 3/4}$ or 3.35. From Table 17 we know that 5% of the frequency in a normal distribution lie above 1.65σ. Multiplying our obtained σ (3.35) by 1.65, we get 5.53; and this value when added to 15 gives us 20.5 as the point above which lie 5% of the "chance" distribution of scores. A score of 21 (20.5 to 21.5), therefore, may be regarded as significant, and if a subject achieves such a score we can be reasonably sure that he is not merely guessing.

For a higher level of assurance, we may take that score which would occur by chance only once in a hundred trials. From Table 17, 1% of the frequency in the normal curve lies above 2.33σ. This point is 7.81 (3.35×2.33) above 15 or at 22.8. A score of 23, therefore, or a higher score is *very* significant; only once in one hundred trials would a subject achieve such a score by guessing.

Use of the normal probability curve in the solution of problems like this always involves a degree of approximation. When p differs considerably from 1/2 and n is small, the distribution resulting from the expansion of $(p + q)^n$ is skewed and is not therefore accurately described by the normal curve. Under these circumstances one must resort to the direct determination of probabilities as in Example (1). When n is large, however, and p not far from 1/2, the normal distribution can be safely used, as will be shown by the chi-square tests on page 245.

II. The χ^2 (Chi-Square) Test

The chi-square test represents a useful method of evaluating experimentally determined results against results to be expected on some hypothesis. The formula for chi-square (χ^2) is

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right] \tag{42}$$

(chi-square formula for testing agreement between observed and expected results)

in which

f_o = frequency of occurrence of observed or experimentally determined facts;

f_e = expected frequency of occurrence on some hypothesis.

The differences between observed and expected frequencies are squared and divided by the expected number in each case, and the sum of these quotients is χ^2. The more closely the observed results approximate to the expected, the smaller is chi-square and the closer the agreement between the observed data and the hypothesis being tested. On the other hand, the larger the chi-square, the greater the probability of a real divergence of experimentally observed results from expected results. To evaluate chi-square, we enter Table 32 with the given value of chi-square and with df, the number of degrees of freedom. The quantity $df = (r - 1)(c - 1)$ in which r is the number of rows and c the number of columns in which the data are tabulated. From Table 32 we find P, the probability that the obtained χ^2 is significant. Several illustrations of the chi-square test will be given in the sections following.

1. Testing the Divergence of Observed Values from Values Calculated on the Hypothesis of Equal Probability (Null Hypothesis)

Example (1) Forty-eight subjects are asked to express their attitude toward the proposition "Should the United States Join a Security Organization of Nations?" by marking F (favorable) I (indifferent) or U (unfavorable). Of the

TABLE 32

Table of Chi-Square

(The values of χ^2 are printed in the body of the table.)

Adapted from R. A. Fisher's *Statistical Method for Research Workers*, Oliver & Boyd, by permission of publishers.

df	P = 0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0.000157	0.000628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21,161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

members in the group, twenty-four marked *F*, twelve *I*, and twelve *U*. Do these results indicate a significant trend of opinion?

The observed data (f_o) are given in the first row of Table 33. In the second row is the distribution of answers to be expected on the null hypothesis (f_e), if each answer is selected equally often. Below the table are entered the differences ($f_o - f_e$). Each of these differences is squared and divided by its f_e ($64/16 + 16/16 + 16/16$) to give $\chi^2 = 6$.

TABLE 33

	Answers			
	Favorable	Indifferent	Unfavorable	
Observed (f_o)	24	12	12	48
Expected (f_e)	16	16	16	48
$f_o - f_e$	8	4	4	
$(f_o - f_e)^2$	64	16	16	
$\frac{(f_o - f_e)^2}{f_e}$	4	1	1	

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right] = 6 \qquad df = 2 \qquad P = .05 \text{ (Table 32)}$$

The degrees of freedom in the table may be readily calculated from the formula $df = (r - 1)(c - 1)$ to be $(3 - 1)(2 - 1)$ or 2. Also, the degrees of freedom may be found directly in the following way: Since we know the row totals to be 48, when two entries are made in a row the third is immediately fixed, is not "free." When the first two entries in row 1 are 24 and 12, for example, the third entry must be 12 to make up 48. Since we also know the sums of the columns, only *one* entry in a column is free, the second being fixed as soon as the first is tabulated. There are, then, *two* degrees of freedom for rows and *one* degree of freedom for columns, and $2 \times 1 = 2$ degrees of freedom for the table.

Entering Table 32 we find in row $df = 2$, a χ^2 of almost 6 (actually, 5.991) in the column headed .05. A *P* of .05 means

that should we repeat this experiment, only once in twenty trials would a χ^2 of 6 (or more) be expected to occur *if* the null hypothesis were true. Our result may be marked "significant at the .05 level," therefore, on the grounds that the divergence of observed from expected results is much too large to be attributed *solely* to sampling fluctuations. We reject the "equal answer" hypothesis and conclude that our group really favors the proposition. In general, we may safely discard a null hypothesis whenever P is .05 or less.

Example (2) The items in an attitude scale are answered by underlining one of the following phrases: Strongly approve, approve, indifferent, disapprove, strongly disapprove. The distribution of answers to an item marked by 100 subjects is shown in Table 34. Do these answers diverge significantly from the distribution to be expected if there are no preferences in the group?

TABLE 34

	Strongly Approve	Approve	Indifferent	Disapprove	Strongly Disapprove	
Observed (f_o)	23	18	24	17	18	100
Expected (f_e)	20	20	20	20	20	100
$(f_o - f_e)$	3	2	4	3	2	
$(f_o - f_e)^2$	9	4	16	9	4	
$\frac{(f_o - f_e)^2}{f_e}$	.45	.20	.80	.45	.20	

$\chi^2 = 2.10$ $df = 4$ P lies between .70 and .80

On the null hypothesis of "equal probability" twenty subjects may be expected to select each of the five possible answers. Squaring the $(f_o - f_e)$, dividing by the expected result (f_e), and summing, we obtain a χ^2 of 2.10. $df = (5 - 1)(2 - 1)$ or 4. From Table 32, reading across from row $df = 4$, we locate a χ^2 of 2.195 in column .70. This χ^2 is nearest to our calculated value of 2.10, which lies between the entries in columns .70 and .80. It is sufficiently accurate to describe P as lying be-

tween .70 and .80 without interpolation. Since this much divergence from the null hypothesis, namely, 2.10 can be expected to occur upon repetition of the experiment in approximately 75% of the trials, χ^2 is clearly *not* significant and we must retain the null hypothesis. There is no conclusive evidence of either a favorable or unfavorable attitude toward this item.

2. Testing Divergence of Observed Values from Values Calculated on the Hypothesis of a Normal Distribution

Our hypothesis may assert that the frequencies of an event which we have observed really follow the normal distribution instead of being equally probable. An example illustrates how this hypothesis may be tested by chi-square.

Example (3) Forty-two salesmen have been classified into five groups — excellent, very good, satisfactory, poor, and very poor — by a consensus of sales managers. Does this distribution of ratings differ significantly from that to be expected if selling ability is normally distributed?

TABLE 35

	Excellent	Very Good	Satisfactory	Poor	Very poor	
Observed (f_o)	6	10	20	4	2	42
Expected (f_e)	1.5	10	19	10	1.5	42
$(f_o - f_e)$	4.5	0	1	6	.5	
$(f_o - f_e)^2$	20.25	0	1	36	.25	
$\frac{(f_o - f_e)^2}{f_e}$	13.50	0	.05	3.60	.17	

$\chi^2 = 17.32$ $df = 4$ P is less than .01

The entries in row 1 give the number of men classified in each of the five categories. In row 2, the entries show how many of the forty-two salesmen may be expected to fall in each category on the hypothesis of a normal distribution. These last entries were found by dividing the baseline of a normal curve (taken to extend over 6σ) into five equal segments of 1.2σ each.

From Table 17, the proportions of the normal distribution to be found in each of these segments are as follows:

		Proportion
Between	$+3.00\sigma$ and 1.80σ	.035
"	1.80σ and $.60\sigma$	.24
"	$.60\sigma$ and $-.60\sigma$	.45
"	$-.60\sigma$ and -1.80σ	.24
"	-1.80σ and -3.00σ	.035
		1.000

These proportions taken as percentages of forty-two have been calculated and are entered in Table 35. The χ^2 in the table is 17.32 and $df = (5-1)(2-1)$ or 4. From Table 32 it is clear that this value of χ^2 lies beyond the limits of the table, hence P is listed simply as less than .01. The discrepancy between observed and expected values is so great that the hypothesis of a normal distribution of selling ability must be rejected. Too many men have been described as excellent, and too few as poor and very poor, to make for agreement with our hypothesis.

3. The Chi-Square Test When Table Entries Are Small

When table entries are large, estimates of probability given by the χ^2-test are usually quite close to those obtained by direct methods. But when table entries are small (say five or less), and especially when the table is 2×2 fold (when the number of degrees of freedom is 1) the chi-square test is subject to considerable error. It is customary in such cases to make a correction — called the correction for continuity.* Reasons for making this correction will be best understood from the examples following.

Example (4) In Example (1), page 234, an observer gave seven correct judgments in ten trials. The probability of a right judgment was 1/2 in each instance, so that the expected number of correct judgments was five. Test our

* Goulden, C. H., *Methods of Statistical Analysis* (1939), pp. 101–110. Snedecor, G. W., *Statistical Methods* (3rd ed., 1940), pp. 169–170.

subject's deviation from the null hypothesis by computing chi-square and compare the P with that found by direct calculation.

TABLE 36

	Right	Wrong	
Observed (f_o)	7	3	10
Expected (f_e)	5	5	10
$(f_o - f_e)$	2	2	
Correction (− .5)	1.5	1.5	
$(f_o - f_e)^2$	2.25	2.25	
$\frac{(f_o - f_e)^2}{f_e}$	.45	.45	

$\chi^2 = .90$
$df = 1$
$P = .356$ (by interpolation in Table 32)
$\frac{1}{2}P = .178$

Calculations in Table 36 follow those of previous tables except for the correction which consists in *subtracting* .5 from each $(f_o - f_e)$ difference. In applying the χ^2-test we assume that adjacent frequencies are connected by a continuous and smooth curve (like the normal curve) and are not discrete numbers. In 2×2 fold tables, however, in which the entries are small the curve is not continuous. Hence, the deviation of 7 from 5 must be written as 1.5 (6.5 − 5) instead of 2 (7 − 5), since 6.5 is the lower limit of 7 in a continuous series. In like manner the deviation of 3 from 5 must be taken from the upper limit of 3, namely, 3.5 (see Fig. 46). Still another change in procedure must be made in order to have the probability obtained from χ^2 agree with the direct determination of probability. P in the χ^2 table gives the probability of 7 or more right answers *and* of 3 or less right answers, i.e., it takes account of both ends of the probability curve. We must take 1/2 of P, therefore, if we want only the probability of 7 or more *right* answers. Note that the $P/2$ of .178 is very close to the P of .172 got by the direct method on page 235. If we repeated our test we should expect a score of 7 or better about seventeen times in 100 trials. It is

clear, therefore, that the obtained score is not significant and does not refute the null hypothesis.

It should be noted that had we omitted the correction for continuity, chi-square would have been 1.60 and $P/2$ (by interpolation in Table 32) .095. It is clear that failure to use the correction causes the probability to be greatly *underestimated* and the significance of our result considerably increased.

When the expected entries in a 2×2 fold table are the same (as in Tables 36, 37) the formula for chi-square may be written in a somewhat shorter form as follows:

$$\chi^2 = \frac{2(f_o - f_e)^2}{f_e} \tag{43}$$

(short formula for χ^2 in 2×2 fold tables when expected frequencies are equal)

Applying formula (43) to Table 36 we get a chi-square of $\frac{2(1.5)^2}{5} = .90$.

Example (*5*) In Example (3), page 238, a subject achieved a score of sixty right on a test of 100 true-false items. From the chi-square test, determine whether this subject was merely guessing. Compare your result with that found on page 238 when the normal curve hypothesis was employed.

TABLE 37

	Right	Wrong	
Observed (f_o)	60	40	100
Expected (f_e)	50	50	100
$(f_o - f_e)$	10	10	
Correction (− .5)	9.5	9.5	
$(f_o - f_e)^2$	90.25	90.25	
$\frac{(f_o - f_e)^2}{f_e}$	1.81	1.81	

$\chi^2 = 3.62$ $\qquad$ $P = .059$
$df = 1$ $\qquad$ $\frac{1}{2}P = .0295$ or .03

Although the cell entries in Table 37 are large, use of the correction for continuity will be found to yield a result in some-

what closer agreement with that found on page 238 than can be obtained without the correction. As shown in Figure 46, the probability of a deviation of 60 or more from 50 is that part of the curve lying above 59.5. In Table 37, the P of .059 gives us the probability of a score of 60 or more *and* of 40 or less. Hence we must take 1/2 of P (i.e., .0295) to give us the probability of a score of 60 or more. Agreement between the probability given by the χ^2-test and by direct calculation (p. 238) is very close. Note that when χ^2 is calculated without the correction, we get a $P/2$ of .024, a slight underestimation. In general, the correction for continuity has little effect when table entries are large (as here). But failure to use the correction even when numbers are fairly large may lead to some underestimation of the probability; hence it is generally wise to use it.

> *Example (6)* In Example (4), page 239, given a multiple-choice test of sixty items (four possible answers to each item) we were required to find what score a subject must achieve in order to demonstrate knowledge of the test material. By use of the normal probability distribution, it was shown that a score of 21 is reasonably significant and a score of 23 highly significant. Can these results be verified by the chi-square test?

In Table 38 an obtained score of 21 is tested against an expected score of 15. In the first line of the table the observed

TABLE 38

	R	W	
f_o	21	39	60
f_e	15	45	60
$(f_o - f_e)$	6	6	
Correction (− .5)	5.5	5.5	
$(f_o - f_e)^2$	30.25	30.25	
$\dfrac{(f_o - f_e)^2}{f_e}$	2.02	.67	

$\chi^2 = 2.69$ $\qquad\qquad P = .10$

$df = 1$ $\qquad\qquad \frac{1}{2}P = .05$

values (f_o) are 21 right and 39 wrong; in the second line, the expected or "guess" values are 15 right and 45 wrong. Making the correction for continuity we obtain a χ^2 of 2.69, a P of .10 and 1/2 P of .05. Only once in twenty trials would we expect a score of 21 or higher to occur if the subject were merely guessing, had no knowledge of the test material. This answer checks the result obtained on page 240.

In Table 39 a score of 23 is tested against the expected score of 15. Making the correction for continuity, we obtain a χ^2 of 5.00 which yields a P of .0275 and 1/2 P of .0138. Again this result closely checks the answer obtained on page 240 by use of the normal probability curve.

TABLE 39

	R	W	
f_o	23	37	60
f_e	15	45	60
$f_o - f_e$	8	8	
Correction (− .5)	7.5	7.5	
$(f_o - f_e)^2$	56.25	56.25	
$\frac{(f_o - f_e)^2}{f_e}$	3.75	1.25	

$\chi_2 = 5.00$ $\quad P = .0275$

$df = 1$ $\quad \frac{1}{2}P = .0138$ or .01

4. The X²-Test When Table Entries Are in Percentages

The chi-square test should not be used with percentage entries unless a correction for size of sample is made. This follows from the fact that in dealing with probability the significance of an event depends upon its *actual* frequency and is not shown by its percentage occurrence. For a penny to fall heads eight times in ten throws is not as significant as for the penny to fall heads eighty times in 100 throws, although the percentage occurrence is the same in both cases. If we write the entries in Table 36 as percentages, we have

	R	W	
f_o	70%	30%	100%
f_e	50%	50%	100%
$(f_o - f_e)$	20%	20%	
Correction* (− 5%)	15%	15%	
$(f_o - f_e)^2$	225	225	

$$\chi^2\% = \frac{2(225)}{50} = 9 \qquad \text{by (43)}$$

$$\chi^2 = 9 \times \frac{10}{100} = .90 \text{ (Table 36)}$$

It is clear that in order to bring χ^2 to its proper value in terms of original numbers we must multiply the "percent" χ^2 by 10/100 to give .90. A χ^2 calculated from percentages must always be multiplied by $N/100$ (N = number of observations) in order to adjust it to the actual frequencies in the given sample.

5. The χ^2-Test of Independence in Contingency Tables

We have seen that χ^2 may be employed to test the agreement between observed results and those expected on some hypothesis. A further useful application of chi-square can be made when we wish to investigate the relationship between traits or attributes which can be classified into two or more categories. The same persons, for example, may be classified as to hair color (light, brown, black, red) and as to eye color (blue, gray, brown), and the correspondence in these attributes noted. Or fathers and sons may be classified with respect to interests or temperament or achievement and the relationship of the attributes in the two groups studied.

Table 40 is a contingency table, i.e., a double entry or two-way table in which the possession by a group of varying degrees of two characteristics is represented. In the tabulation in Table 40, 413 persons have been classified as to "eyedness" and "handedness." Eyedness, or eye dominance, is described as

* The unit here is 10%, so that 5% must be subtracted from each $(f_o - f_e)$ difference. Thus (70% − 50%) is actually (65% − 50%), and (30% − 50%) is (35% − 50%). See page 247.

left-eyed, ambiocular, or right-eyed; handedness as left-handed, ambidextrous, or right-handed. Reading down the first column we find that of 118 left-eyed persons, 34 are left-handed, 27 ambidextrous and 57 right-handed. Across the first row we find 124 left-handed persons, of whom 34 are left-eyed, 62 ambiocular and 28 right-eyed. The other columns and rows are interpreted in the same way.

TABLE 40

COMPARISON OF EYEDNESS AND HANDEDNESS IN 413 PERSONS*

	Left-Eyed	Ambiocular	Right-Eyed	Totals
Left-handed	(35.4) 34	(58.5) 62	(30.0) 28	124
Ambidextrous	(21.4) 27	(35.4) 28	(18.2) 20	75
Right-handed	(61.1) 57	(101.0) 105	(51.8) 52	214
Totals	118	195	100	413

I. Calculation of independence values (f_e):

$$\frac{118 \times 124}{413} = 35.4 \qquad \frac{195 \times 124}{413} = 58.5 \qquad \frac{100 \times 124}{413} = 30.0$$

$$\frac{118 \times 75}{413} = 21.4 \qquad \frac{195 \times 75}{413} = 35.4 \qquad \frac{100 \times 75}{413} = 18.2$$

$$\frac{118 \times 214}{413} = 61.1 \qquad \frac{195 \times 214}{413} = 101.0 \qquad \frac{100 \times 214}{413} = 51.8$$

II. Calculation of χ^2:

$(-1.4)^2 \div 35.4 = .055 \quad (3.5)^2 \div 58.5 = .209 \quad (-2.0)^2 \div 30 = .133$

$(5.6)^2 \div 21.4 = 1.465 \quad (-7.4)^2 \div 35.4 = 1.547 \quad (1.8)^2 \div 18.2 = .178$

$(-4.1)^2 \div 61.1 = .275 \quad (4.0)^2 \div 101.0 = .158 \quad (.20)^2 \div 51.8 = .001$

$\chi^2 = 4.02 \quad df = 4 \quad P$ lies between .30 and .50

* From Woo, T. L., *Biometrika* (1936), 20A, pp. 79–118.

The hypothesis to be tested is the null hypothesis, namely, that handedness and eyedness are essentially unrelated or independent. In order to compute χ^2 we must first calculate an "independence value" for each cell in the contingency table. Independence values are represented by figures in parentheses

within the different cells; they give the number of people whom we should expect to find possessing the designated eyedness and handedness combinations in the absence of any real association. The method of calculating independence values is shown in Table 40. To illustrate with the first entry, there are 118 left-eyed and 124 left-handed persons. If there were no association between left-eyedness and left-handedness we should expect to find, by chance, $\frac{118 \times 124}{413}$ or 35.4 individuals in our group who are left-eyed *and* left-handed. The reason for this may readily be seen. We know that 118/413 of the entire group is left-eyed. This proportion of left-eyed individuals should hold for any sub-group, if there is *no* dependence of eyedness on handedness. Hence, 118/413 or 28.5% of the 124 left-handed individuals, i.e., 35.4, should also be left-eyed. Independence values for all cells are shown in Table 40.

When the expected or independence values have been computed, we find the difference between the observed and expected values for each cell, square each difference and divide in each instance by the independence value. The sum of these quotients by formula (42) gives χ^2. In the present problem $\chi^2 = 4.02$ and $df = (3 - 1)(3 - 1)$ or 4. From Table 32 we find that P lies between .30 and .50 and hence χ^2 is not significant. The observed results are close to those to be expected on the hypothesis of independence and there is no evidence of any real association between eyedness and handedness within our group.

III. The Analysis of Variance

Analysis of variance represents still another means of testing the null hypothesis. The term "analysis of variance" includes (*a*) a variety of experimental designs or arrangements, as well as (*b*) certain statistical techniques appropriate for use with these designs. The statistical methods employed in analysis of variance are not new (as they are often thought to be), but are, in reality, adaptations of methods described earlier in this book. The experimental designs, on the other hand, are in many

instances new — at least to psychology. These systematic procedures will often provide a more efficient test of the null hypothesis than methods now customarily used.

In the following sections certain elementary applications of analysis of variance to experimental psychology will be shown by means of a problem which illustrates the simplest design. It is hoped that by working through this problem the reader will become acquainted with the mechanics of analysis of variance, as well as with some of its possibilities. For further and more comprehensive treatments of this topic the reader should consult the books listed below.* Only a brief outline is attempted here.

1. How Variances Can Be Analyzed

The variability within a set of scores (N large) may be measured by the standard deviation $\left(\sqrt{\frac{\Sigma x^2}{N}}\right)$, but it may also be expressed in terms of the "variance" or $\sigma^2 \left(\frac{\Sigma x^2}{N}\right)$. A decided advantage of variances over SD's is that variances are oftentimes additive — and the sums of squares upon which variances are based always are. As an example, suppose we add the two independent scores X and Y to get the composite score Z. Expressing x, y, and z as deviations from their means, M_x, M_y, and M_z, we may write

$$z = x + y$$

and squaring and summing, $\Sigma z^2 = \Sigma x^2 + \Sigma y^2$. (The term in xy drops out since there is no correlation between x and y — x and y are independent by hypothesis.) Dividing by N, we have

* Snedecor, G. W., *Statistical Methods* (1946), Chapters 10, 11, 12, 13, 15, and 17.

Goulden, C. H., *Methods of Statistical Analysis* (1939), Chapters 5, 11, 13, and 15.

Lindquist, E. F., *Statistical Analysis in Educational Research* (1940), Chapters 4, 5, and 6.

Fisher, R. A., *The Design of Experiments* (1935).

Fisher, R. A., *Statistical Methods for Research Workers* (8th ed., 1941). (The Fisher references will be difficult for the beginner.)

$$\sigma^2_z = \sigma^2_x + \sigma^2_y$$

and
$$\sigma_z = \sqrt{\sigma^2_x + \sigma^2_y}$$

The first equation in terms of variances is more convenient for analysis than is the equation in terms of standard deviations. If we divide through by σ_z^2, for example, we find that $1 = \frac{\sigma^2_x}{\sigma^2_z} + \frac{\sigma^2_y}{\sigma^2_z}$; from which we are able to determine what *proportion* of the total variance (σ^2_z) is attributable to the variance of X and what proportion is attributable to the variance of Y. Analysis into proportional contributions cannot be made with standard deviations.

The technique of variance analysis is illustrated by the data in Table 41. From a large group of fifth-grade boys, four boys are given a test under condition A, four under condition B, and four under condition C. Subjects are assigned at random to each of the three groups. Do the mean scores achieved under conditions, A, B, and C differ significantly?

We may begin with the null hypothesis, namely, that the three different conditions do not really influence the final scores and that variations in the performance of the three groups are no greater than might be expected by chance. To test this hypothesis we may compare the variation attributable to the different methods with the variation to be expected in a group of boys all of whom have taken the test under the same method. The variation exhibited by all twelve boys is to be divided, then, into two portions: (1) the variance attributable to methods (the between-methods effect), and (2) the variance attributable to subjects (the within-groups effect), and these two variances are to be compared. The procedure is outlined in the following steps which parallel the calculations in Table 41:

Step 1

The total variation is obtained first by summing up the squares of the deviations from the mean of all twelve boys. The general mean is 12; and the sum of the squares of deviations around this general mean (GM) is 178.

TABLE 41

To Illustrate the Use of the Ratio "Variance among Methods" to "Variance within Groups" in Determining the Significance of Differences among Means. Three Groups of Four Subjects Each Work by Methods A, B, and C. The Data Are Artificial.

	Methods		
	A	B	C
	10	15	10
	14	20	12
	12	17	6
	8	8	12
	44	60	40
M's =	11	15	10

General Mean (GM) $= \frac{11 + 15 + 10}{3} = 12$

Steps:

1. Sum of squares of deviations of the scores in A, B, and C around GM of $12 = 2^2 + 2^2 + 0 + 4^2 + 3^2 + 8^2 + 5^2 + 4^2 + 2^2 + 0 + 6^2 + 0 = 178$

2. Sum of squares of deviations of M's of A, B, and C around GM of $12 = (11 - 12)^2 + (15 - 12)^2 + (10 - 12)^2 = 1^2 + 3^2 + 2^2 = 14$

3. (*a*) For list A: Sum of squares of deviations around M of $11 = 1^2 + 3^2 + 1^2 + 3^2 = 20$

   (*b*) For list B: Sum of squares of deviations around M of $15 = 0 + 5^2 + 2^2 + 7^2 = 78$

   (*c*) For list C: Sum of squares of deviations around M of $10 = 0 + 2^2 + 4^2 + 2^2 = 24$

   Total sum of squares of deviations *within* groups $= 122$

4. $178 = 122 + 56$ (i.e., 4×14). (4 = number of cases in a given column)

5. Degrees of freedom for sum of squares of deviations around GM = degrees of freedom for sum of squares of deviations within groups + degrees of freedom for sum of squares of deviations among methods; or $11 = 9 + 2$.

6. Variance (between methods) $= \frac{56}{2} = 28$.

   Variance (within groups) $= \frac{122}{9} = 13.56$.

   F-test: $F = \frac{\text{variance (between)}}{\text{variance (within)}} = \frac{28}{13.56} = 2.07.$

The F of 2.07 is smaller than 4.26 and is not significant at the .05 level.

Step 2

The between-methods variance is obtained in the following way: The mean of method A is 11; of method B, 15; and of method C, 10. The sum of the squares of the deviations of these three means (11, 15, and 10) around the GM of 12 is 14.

Step 3

The variance attributable to subjects, sometimes called the residual variance, is found by adding the sums of squares within columns. The sum of squares of deviations of the four scores in A around their mean of 11 is 20; the sum of squares of deviations of the four scores in B around their mean of 15 is 78; and the sum of squares of deviations of the four scores in C around their mean of 10 is 24. Adding these, we get 122 as the sum of squares of deviations *within* the columns A, B, C. The sum of squares in each column is around its own mean. Hence the final sum gives the variation attributable to subjects, and is independent of systematic differences from column to column.

Step 4

Writing the sums of squares in the form of an equation, we have that $178 = 122 + 4 \times 14$, or sum of squares around GM = sum of squares *within* methods + n (i.e., 4) $\times$ sum of squares *between* the M's of methods. The sum of squares around a GM can always be broken down (as here) into component sums of squares.

2. Degrees of Freedom

Each of these sums of squares becomes a variance when divided by the appropriate number of degrees of freedom.

Step 5

Since there are 12 scores in all (A + B + C), the divisor for 178 (sum of squares of deviations around GM) is $(N - 1)$ or 11 degrees of freedom. The divisor for 122 (sum of squares of deviations around the group means) is 9 degrees of freedom, as there are $(n - 1)$ or 3 degrees of freedom in each list and 3×3 or 9 degrees of freedom in the three lists. This leaves 2 degrees of freedom as the divisor for 4×14 (sum of squares of deviations among M's). Expressing the degrees of freedom as an equation, we have $11 = 9 + 2$, or degrees of freedom for sum of squares of *deviations* around GM = degrees of freedom *within* groups *plus* degrees of freedom for sum of squares of deviations among M's of methods.

3. Measuring Significance by Means of the Ratio of "Between" to "Within" Variance

Step 6

Dividing 122 by 9, we get 13.56 as the variance *within* our three groups; and dividing 56 by 2, 28 as the variance among the means of our three methods. In this problem the null hypothesis asserts that the three sets of scores A, B, and C are random samples drawn from the same parent population and that their M's differ only through sampling accidents. This hypothesis may be tested by computing the ratio "F." $F = \frac{\text{between } M\text{'s variance}}{\text{within group variance}}$; or in our problem, $\frac{28}{13.56} = 2.07$. The significance of F depends upon the degrees of freedom in the numerator and in the denominator of the fraction which determines F. From tables of F,* we find that when the numerator has 2 degrees of freedom and the denominator 9, F must equal 4.26 to be significant at the .05 level of confidence and 8.02 to be significant at the .01 level of confidence.

Our F falls far below the .05 level, hence there is no assurance of any actual differences among our method means. We retain the null hypothesis, since on the present evidence there is no reason to believe our groups to be other than random samples drawn from the same population.

In the next section, another problem similar to that of Table 41 is given to illustrate the procedure usually followed in analysis of variance. The data in Table 42 constitute a simple but fundamental experimental design which is often useful.

4. An Illustration of Simple Analysis of Variance When There Is One Criterion of Classification

Example (1) A sensory-motor learning test is administered to groups of subjects under five conditions or methods, designated, respectively, A, B, C, D, and E. Five subjects are

* For F-tables see Snedecor, *op. cit.*, pp. 184–187; or Lindquist, *op. cit.*, pp. 62–65.

assigned at random to each group. Do the mean scores achieved under the five methods differ significantly?

Records for each of the five groups are shown in parallel columns in Table 42. Individual scores are listed under the five headings which designate the conditions under which the learning test was administered. Since "methods" furnishes the only categories, there is said to be *one* criterion of classification. The first object of our analysis is a breakdown of the *total variance* (σ^2) of the twenty-five scores into two parts: (1) the variance attributable to methods, and (2) the variance attributable to individual differences, i.e., within the several groups. Computation of the sums of squares upon which these variances are based is shown in Table 42 A. A more detailed account of these calculations may be set forth as follows:

Step 1

Calculation of the "correction term." When the SD is calculated from original measures,* the formula $\sigma^2 = \frac{\Sigma x^2}{N} - c^2$ becomes $\sigma^2 = \frac{\Sigma X^2}{N} - M^2$. The correction equals the mean (M) directly since $AM = 0$. Replacing σ^2 by $\frac{\Sigma x^2}{N}$, we have that $\frac{\Sigma x^2}{N} = \frac{\Sigma X^2}{N} - M^2$. If the correction term M^2 is written $\frac{(\Sigma X)^2}{N^2}$, we may multiply this equation through by N to find that $\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$. In Table 42 the correction term $\frac{(\Sigma X)^2}{N}$ is $\frac{(1135)^2}{25}$ or 51,529.0.

Step 2

Since $\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$, we must square and sum the original scores and then subtract the correction term (51,529),

* See page 62. It is customary in analysis of variance to calculate variances from original measures or scores.

TABLE 42

SCORES MADE BY FIVE GROUPS OF STUDENTS ON A LEARNING TEST

Each group consists of five individuals and each group takes the test by a different method. To illustrate analysis of variance when there is *one* criterion of classification:

	Methods					
	A	B	C	D	E	
	35	38	34	55	71	
	26	50	26	65	59	
	29	50	59	56	43	
	37	36	23	71	63	
	40	40	60	35	34	
Sums =	167	214	202	282	270	1135
M's =	33.4	42.8	40.4	56.4	54.0	GM = 45.4

A. *Calculation of Sums of Squares (Computation from Original Measure)*

Step 1. Correction term (C) $= \frac{(1135)^2}{25} = \frac{1288225}{25} = 51529.0$

Step 2. Total sum of squares

$$= (35^2 + 26^2 + 29^2 + \ldots + 34^2) - C$$
$$= 56641 - 51529 = 5112$$

Step 3. Sum of squares among means of methods A, B, C, D, and E

$$= \frac{(167)^2 + (214)^2 + (202)^2 + (282)^2 + (270)^2}{5} - C$$
$$= 53382.6 - 51529.0 = 1853.6$$

Step 4. Sum of squares *within* methods $= 5112 - 1853.6 = 3258.4$

B. *Analysis of Variance*

Source	*df*	Sum of Squares	Mean Sq. (Variance)	*SD*
Among the means of methods	4	1853.6	463.4	
Within methods	20	3258.4	162.9	12.8
Total	24	5112.0		

$$F = \frac{463.4}{162.9} = 2.84$$

From Table (For 4/20 *df*)

F at .05 = 2.87
F at .01 = 4.43

in order to find the sum of squares around the mean of all twenty-five scores. In Table 42, squaring each score and summing, we get a total of 56,641; and subtracting the correction, the final result is 5,112. This sum of squares can also be computed from the deviations around the means. The general mean is 45.4; subtracting 45.4 from each of the 25 scores,

squaring these deviations and summing, we get 5112, which checks the above.

Step 3

To find the sum of squares attributable to methods we square the sum of each column, add these values, and divide the total by five (the number of individuals in each column). If now we subtract the correction found in Step 1, the resulting sum of squares is 1853.6. As we are still working with original measures, the method of calculation here repeats Step 1, except that we must divide the sum of squares for column totals by the number of scores in each column.

Step 4

The sum of squares within columns (individual variation) always equals the total sum of squares minus the sum of squares among the means of columns. Our within columns sum is found by subtracting 1853.6 from 5112 to give 3258.4. It may also be calculated directly from the data.*

Calculation of the variances from the three sums of squares, and the analysis of the total variance in terms of its two components is shown in Table 42 B. Each sum of squares must be divided by the number of degrees of freedom allotted to it in order to give the mean square or variance shown in the fourth column under "B." There are twenty-five scores in all in Table 42 and $(N - 1)$ or 24 degrees of freedom. The degrees of freedom for methods are listed as $(5 - 1)$ or 4, less by 1 than the number of methods; and the degrees of freedom within columns are $(24 - 4)$ or 20. This last *df* may be calculated directly in the following way: there are $(5 - 1)$ or 4 degrees of freedom in each column; and 4×5 (number of methods or columns) gives us 20 degrees of freedom for within groups.

The significance of the differences among the means of our five methods can be determined by dividing methods variance by within groups variance to give the ratio called F. From

* For an illustration, see Goulden, *op. cit.*, Example 29, pp. 125–127.

tables of F * we find that an F of 2.87 † represents the ratio which, under the conditions of our problem, is significant at the .05 level; and an F of 4.43 † represents the ratio which is significant at the .01 level. Since our calculated F of 2.84 is almost equal to 2.87, we may regard the null hypothesis as barely disproved at the .05 level of confidence. According to the evidence, therefore, the five methods differ significantly.

F furnishes a comprehensive or over-all test of significance. A significant F does not tell us *which* method is best but simply that one or more differences as between method-means must be significant. If F is not significant there is no point in going further as *no* mean difference can be significant. But if F is significant we may proceed to calculate CR's for the differences between column means by the following method.

Means for the five methods are given in Table 42; they vary from 33.4 to 56.4. The best estimate of experimental or individual variability is given by the SD computed from the within groups variance in "B" of Table 42. This SD is based upon *all* of our data and gives the variability in the table *after* the systematic effect of methods-differences has been removed. (Note analogy here to "partial" σ, p. 417.) Hence, it is used instead of the SD's calculated from the separate columns, A, B, C, D, and E. The SE of any mean (SE_M) will be $\frac{SD}{\sqrt{N}}$ or $\frac{12.8}{\sqrt{5}} = 5.7$; and the SE of any mean difference is (SE_D)

$$\sqrt{SE^2_{M_1} + SE^2_{M_2}} \text{ or } \sqrt{\frac{(12.8)^2}{5} + \frac{(12.8)^2}{5}} = 8.1.$$

Instead of dividing each difference between M's by 8.1 and evaluating the significance of the resulting CR (or t) we may calculate directly that difference which is significant at the .05 level, and check the obtained differences against it. From Table 29 we know that for 20 degrees of freedom a t of 2.09 is

* For F-tables see Snedecor, *op. cit.*, pp. 184–187; or Lindquist, *op. cit.*, pp. 62–65.

† For 4/20 degrees of freedom.

significant at the .05 level. Hence, since $D = t \times SE_D$, we find, upon substituting 2.09 for t and 8.1 for SE_D, that a difference of 16.9 is significant at the .05 confidence level. Table 43 below gives D's between pairs of M's, and the significance of these differences.

TABLE 43

Methods	Differences	Significant at .05 level
A–B	− 9.4	no
A–C	− 7.0	no
A–D	− 23.0	yes
A–E	− 20.6	yes
B–C	2.4	no
B–D	− 13.6	no
B–E	− 11.2	no
C–D	− 16.0	no (?)
C–E	− 13.6	no
D–E	2.4	no

Both methods D and E are significantly better than A and considerably better than B and C. But methods D and E are not distinguishably different.

Several additional comments may serve to summarize the steps in the solution of our problem in Table 42:

(1) First, it must be remembered that we are testing the null hypothesis — the hypothesis that there are no differences among method-means. Stated in another way, we are testing the hypothesis that our five groups are in reality random samples drawn from the same normally distributed parent population. The F-test refutes the null hypothesis by demonstrating differences among means which would not arise more than once in twenty trials if the null hypothesis were true. Hence, F is significant at the .05 level of confidence; and our groups cannot be random samples from the same population. The t-test tells us which differences are significant.

(2) The 24 degrees of freedom (1 less than 25, the total number of scores) are broken down into 4 degrees of freedom allotted

to the five methods and 20 degrees of freedom allotted to individual variations (within column variance).

(3) According to the traditional method of treating a problem of this sort, standard deviations around the means of the five scores in each column are first computed. From these *SD*'s standard errors of the means and standard errors of the differences among means are found. *CR*'s (or *t*'s) are then calculated for the differences between pairs of means and their significances determined from Table 29. Instead of following this procedure, we have computed in Table 42 a *single SD* based upon the variability within *all* five columns. This is a better estimate of the experimental variation within the table than could be found from the five separate *SD*'s, each based upon five scores. Moreover, it represents variability from which systematic method differences have been removed. Justification for pooling scores lies in our original assumption that under the null hypothesis the five groups are random samples from the same population. To be sure, the *F*-test later disproves this hypothesis; but we may proceed on it as our best assumption until it is — or is not — disproved.

PROBLEMS

1. Two sharp clicking sounds are presented in succession, the second being always more intense or less intense than the first. Presentation is in random order. In eight trials an observer is right six times. Is this result significant?
   (*a*) Calculate *P* directly (p. 234).
   (*b*) Check *P* found in (*a*) by χ^2-test (p. 246). Compare *P*'s found with and without correction for continuity.
2. A multiple-choice test of fifty items provides five responses to each item. How many items must a subject answer correctly
   (*a*) to reach the .05 confidence level?
   (*b*) to reach the .01 confidence level?
3. A multiple-choice test of thirty items provides three responses for each item. How many items must a subject answer correctly before the chances are only one in fifty that he is merely guessing?

4. A pack of fifty-two playing cards contains four suits (diamonds, clubs, spades, and hearts). A subject "guesses" through the pack of cards, naming only suits, and is right eighteen times.
   (*a*) Is this result better than "chance"? (Hint: In using the probability curve compute area to 17.5, lower limit of 18.0, rather than to 18.0.)
   (*b*) Check your answer by the χ^2-test (p. 246).
5. Twelve samples of handwriting, six from normal and six from insane adults, are presented to a graphologist who claims he can identify the writing of the insane. How many "insane" specimens must he recognize correctly in order to prove his contention?
6. The following judgments were classified into six categories taken to represent a continuum of opinion:

	Categories						
	I	II	III	IV	V	VI	Total
Judgments:	8	21	42	51	17	5	144

   (*a*) Test given distribution versus "equal probability" hypothesis.
   (*b*) Test given distribution versus normal distribution hypothesis.
7. In 120 throws of a single die, the following distribution of faces was obtained:

	Faces						
	1	2	3	4	5	6	Total
Observed frequencies:	30	25	18	10	22	15	120

   Do these results constitute a refutation of the "equal probability" (null) hypothesis?
8. The following table represents the number of boys and the number of girls who chose each of the five possible answers to an item in an attitude scale.

	Strongly Approve	Approve	Indifferent	Disapprove	Strongly Disapprove	Total
Boys	25	30	10	25	10	100
Girls	10	15	5	15	15	60

   Do these data indicate a significant sex difference in attitude toward this question? (Note: Test the "independence (null) hypothesis.")

9. The table below shows the number of normals and abnormals who chose each of the three possible answers to an item on a neurotic questionnaire.

	Yes	No	?	Total
Normals	14	66	10	90
Abnormals	27	66	7	100
	41	132	17	190

Does this item differentiate between the two groups? Test the independence hypothesis.

10. From the table below, determine whether Item 27 differentiates between two groups of high and low general ability.

Numbers of Two Groups Differing in General Ability Who Pass Item 27 in a Test

	Passed	Failed	Total
High Ability	31	19	50
Low Ability	24	26	50
	55	45	100

11. The following four sets of measurements were made at different times under the same conditions. Do they differ significantly? Apply the method of analysis of variance given on page 260. The F ratios for 3/16 degrees of freedom are, at the .05 level, 3.24; at the .01 level, 5.29.

Set I	Set II	Set III	Set IV
16	19	17	14
18	19	18	18
20	20	14	12
20	25	16	18
17	25	12	16
٩١	١٠٩	٧٧	٧٨

ANSWERS

1. (*a*) $P = .145$ not significant
   (*b*) $P = .145$ when corrected; .085. uncorrected
2. (*a*) 15
   (*b*) 17
3. 15

4. Probability of 18 or better is .08; not significant
5. 5 or 6 (Probability of 5 or 6 = 37/924 = .04)
6. (*a*) $\chi^2 = 72$; P less than .01 and hypothesis of "equal probability" must be discarded.
   (*b*) $\chi^2 = 11.24$; P is less than .05, and the deviation from the normal hypothesis is significant.
7. Yes. $\chi^2 = 12.90$, $df = 5$, and P is between .02 and .05.
8. No. $\chi^2 = 7.03$, $df = 4$, and P is between .20 and .10
9. No. $\chi^2 = 4.14$, $df = 2$, and P is between .20 and .10
10. No. $\chi^2 = 1.98$, $df = 1$, and P lies between .20 and .10
11. Yes. $F = 6.55$; significant at .01 level

CHAPTER IX

LINEAR CORRELATION

I. The Meaning of Correlation

1. Correlation as a Measure of Relationship

In previous chapters we have been concerned with methods of computing statistical measures designed to represent in a reliable way the performance of an individual or a group in some defined capacity or trait. Frequently, however, it is of more importance to examine the *relationship* of one ability to another than it is to measure performance in either trait alone. Are certain abilities closely related, and others relatively independent? Is it true that good pitch discrimination accompanies musical achievement; or that bright children tend to be less neurotic than average children? If we know the general intelligence of a child, as measured by a standard test, can we say anything about his probable scholastic achievement as represented by grades? Problems like these and many others which involve the relations among abilities are studied by the method of correlation.

When the relationship between two sets of measures is "linear," i.e., can be described by a straight line,* the correlation between the scores may be expressed by the "product-moment" coefficient of correlation. This coefficient is designated by the letter r. The method of calculating r will be outlined in Section III. Before taking up the details of calculation, we shall try to make clear what correlation means, and how r measures relationship.

Let us consider, first, a situation in which relationship is fixed and unchanging. The circumference of a circle is always 3.1416

* See pages 309–311 for a further discussion of "linear" relationship.

times its diameter ($C = 3.1416D$), and this equation holds no matter how large or how small the circle, or in what part of the world we find it. Each time the diameter of a circle is increased or decreased, the circumference is increased or decreased by just 3.1416 times the same amount. In short, the dependence of circumference upon diameter is complete; hence, the correlation between the two dimensions is said to be perfect, and $r = 1.00$. In the same fashion, the relationship between two abilities, as represented by two sets of scores, may also be perfect. Suppose, for example, that a hundred students have exactly the same standing in two tests: — the student who ranks first in the one test ranks first in the other, the student who ranks second in the first test ranks second in the other, and this one-to-one correspondence holds throughout the entire list. The relationship here is perfect since the relative position of each subject is exactly the same in one test as in the other. The coefficient of correlation is 1.00.

Now let us consider the case in which there is *no* correlation present. Suppose that we have administered to one hundred college seniors the Army Alpha Examination and a simple "tapping test" in which the number of separate taps made in thirty seconds is recorded. Let the mean Alpha score for the whole group be 175, and the mean tapping rate be 185 taps in thirty seconds. Now suppose that when we divide our group into three sub-groups in accordance with the size of their Alpha scores, we find that the mean tapping rate of the superior or "high" group (whose mean Alpha score is 190) is 184 taps in thirty seconds; the mean tapping rate of the "middle" group (whose mean Alpha score is 175) is 186 taps in thirty seconds; and the mean tapping rate of the "low" group (whose mean Alpha score is 160) is 185 taps in thirty seconds. Since the tapping rate is almost identically the same for all three groups, it is clear that from a student's tapping rate alone we should be unable to draw any conclusion as to his probable performance upon Alpha. A tapping rate of 185 is as likely to be found with an Army Alpha score of 150, as with one of 175 or even 200.

In other words, there is no correspondence between the scores made by the members of our group upon the two tests, and hence, r, the coefficient of correlation, is zero.*

Perfect relationship, then, is expressed by a coefficient of 1.00, and just no relationship by a coefficient of .00. Between these two limits, varying degrees of relation are indicated by such coefficients as .33, or .65, or .92. A coefficient of correlation falling between .00 and 1.00 always implies *some* degree of positive association, the degree of the association depending upon the size of the coefficient.

Relationship may be negative as well as positive; that is, a *high* degree of one trait may be associated with a *low* degree of another. When negative or inverse relationship is perfect, $r = -1.00$. To illustrate, suppose that in a small class of ten schoolboys, the boy who stands first in Latin ranks lowest (tenth) in shop work; the boy who stands second in Latin ranks next to the bottom (ninth) in shop work; and that each boy stands just as far from the top of the list in Latin as from the bottom of the list in shop work. Here the correspondence between achievement in Latin and performance in shop work is one-to-one and definite enough, but the *direction* of the relationship is inverse and $r = -1.00$. Negative coefficients may range from -1.00 up to .00, just as positive coefficients may range from .00 up to 1.00. Coefficients of $-.20$, $-.50$, or $-.80$ indicate increasing degrees of negative or inverse relationship, just as positive coefficients of .20, .50, and .80 indicate increasing degrees of positive relationship.

2. Correlation Expressed as Agreement between Ranks

The notion underlying correlation can often be most readily comprehended from a simple graphic treatment. Three examples will be given to illustrate values of r of 1 00, -1.00, and approximately .00. Correlation is rarely computed when the

* It may be noted that the *number* of groups (here 3) is unimportant: any convenient set may be used. The important point is that when the correlation is zero, one cannot predict a person's score on the second test knowing his score on the first test.

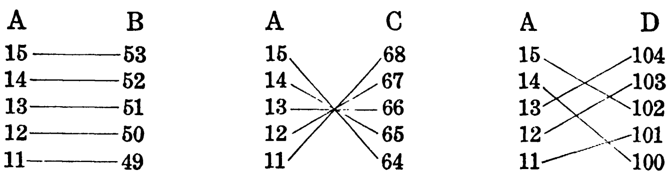
number of cases is less than twenty-five, so that the examples here presented must be considered to have illustrative value only.

Suppose that four tests, A, B, C, and D, have been administered to a group of five children. The children have been arranged in order of merit on Test A and their scores are then compared separately with Tests B, C, and D to give the following three cases:

Case 1			Case 2			Case 3		
Pupil	A	B	Pupil	A	C	Pupil	A	D
a	15	53	a	15	64	a	15	102
b	14	52	b	14	65	b	14	100
c	13	51	c	13	66	c	13	104
d	12	50	d	12	67	d	12	103
e	11	49	e	11	68	e	11	101

Now if the *second* series of scores under each case (i.e., B, C, and D) is arranged in order of merit from the highest score down, and the two scores earned by each child are connected by a straight line, we have the following graphs:

Case 1 Case 2 Case 3



All connecting lines are horizontal and parallel, and the correlation is positive and perfect. $r = 1.00$

All connecting lines intersect in one point. The correlation is negative and perfect, and $r = -1.00$

No system is exhibited by the connecting lines, but the resemblance is closer to Case 2 than to Case 1. Correlation low and negative

The more nearly the lines connecting the paired scores are horizontal and parallel, the higher the positive correlation. The more nearly the connecting lines tend to intersect in one point, the larger the negative correlation. When the connecting lines show no systematic trend, the correlation approaches zero.

3. Summary on Correlation

To summarize our discussion up to this point, coefficients of correlation range over a scale which extends from − 1.00 through .00 to 1.00. A positive correlation indicates that *large* amounts of the one variable tend to accompany *large* amounts of the other; a negative correlation indicates that *small* amounts of the one variable tend to accompany *large* amounts of the other. A zero correlation indicates no consistent relationship. We have illustrated above only perfect positive, perfect negative, and approximately zero correlation in order to bring out the meaning of correlation in a striking way. Only rarely, if ever, however, will a coefficient fall at either extreme of the scale, i.e., at 1.00 or − 1.00. In most actual problems, calculated *r*'s fall at intermediate points, such as .72, − .26, .50, etc. Such *r*'s are to be interpreted as "high" or "low" depending in general upon how close they are to ± 1.00. Interpretation of the degree of relationship expressed by *r* in terms of various criteria will be discussed later on pages 333–339.

II. The Coefficient of Correlation *

1. The Coefficient of Correlation as a Ratio

The product-moment coefficient of correlation may be thought of essentially as that *ratio* which expresses the extent to which changes in one variable are accompanied by — or are dependent upon — changes in a second variable. As an illustration, consider the following simple example which gives the paired heights and weights of five college seniors:

* This section may be taken up after Section III.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Student	Ht. in inches	Wt. in lbs.						
	X	Y	x	y	xy	$\frac{x}{\sigma_x}$	$\frac{y}{\sigma_y}$	$\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}\right)$
a	72	170	3	0	0	1.34	.00	.00
b	69	165	0	− 5	0	.00	− .37	.00
c	66	150	− 3	− 20	60	− 1.34	− 1.46	1.96
d	70	180	1	10	10	.44	.73	.32
e	68	185	− 1	15	− 15	− .44	1.10	− .48
					55			1.80

$M_X = 69$ in. $\sigma_x = 2.24$*
$M_Y = 170$ lbs. $\sigma_y = 13.69$ lbs.* correlation $= \dfrac{\Sigma\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}\right)}{N} = \dfrac{1.80}{5} = .36$

From the X and Y columns it is evident that tall students tend to be somewhat heavier than short students, and hence the correlation between height and weight is almost certainly positive. The mean height is 69 inches, the mean weight 170 pounds, and the σ's are 2.24 inches and 13.69 pounds, respectively. In column (4) are given the deviations (x's) of each man's height from the mean height, and in column (5) the deviations (y's) of each man's weight from the mean weight. The product of these paired deviations (xy's) is a measure of the agreement between individual heights and weights, and the larger the sum of the xy column the higher the degree of correspondence. When agreement is perfect (and $r = 1.00$) the Σxy column has its maximum value. It may be surmised — and with much reason — that the sum of the xy's divided by N (i.e., $\frac{55}{5} = 11$) should give a suitable measure of the relationship between X and Y. Such an average is *not* a stable measure of relationship, however, as it depends directly upon the *units* in which height and weight have been expressed, and consequently will vary (as shown in the example below) if centimeters and kilograms, say, are employed instead of inches and pounds. One may avoid the troublesome matter of differences in units by dividing each x

* These σ's were calculated by formula $\left(s = \sqrt{\frac{\Sigma x^2}{N-1}}\right)$ since the samples are small (see p. 189).

and each y by its own σ, i.e., by expressing each deviation as a standard or z-score. The sum of the products of the standard scores — column (9) — divided by N will then yield a ratio which, as we shall see later, *is* a stable expression of relationship. This ratio is the "product-moment" * coefficient of correlation. Its value of .36 indicates a fairly good positive correlation between height and weight in this small sample. The reader should note that our ratio or coefficient is simply the *average product* of the standard scores of corresponding X and Y measures.

Let us now investigate the effect upon our ratio of changing the units in terms of which X and Y have been expressed. In the example below, the heights and weights of the same five students are expressed (to the nearest whole number) in centimeters and kilograms instead of in inches and pounds.:

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Student	Ht. in cms.	Wt. in kgs.						
	X	Y	x	y	xy	$\frac{x}{\sigma_x}$	$\frac{y}{\sigma_y}$	$\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}\right)$
a	183	77	8	0	0	1.43	.00	.00
b	175	75	0	− 2	0	.00	− .32	.00
c	168	68	− 7	− 9	63	− 1.25	− 1.43	1.79
d	178	82	3	5	15	.53	.80	.42
e	173	84	− 2	7	− 14	− .36	1.11	− .40
					64			1.81

$M_X = 175$ cms. $\sigma_x = 5.61$ cms.†
$M_Y = 77$ kgs. $\sigma_y = 6.30$ kgs.† correlation $= \dfrac{\Sigma\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}\right)}{N} = \dfrac{1.81}{5} = .36$

The mean height of our group is now 175 cms. and the mean weight 77 kgs.; the σ's are 5.61 cms. and 6.30 kgs., respectively.

* The sum of the deviations from the mean (raised to some power) and divided by N is called a "moment." When pairs of deviations in x and y are multiplied together, summed, and divided by N $\left(\text{to give } \frac{\Sigma xy}{N}\right)$ the term "product-moment" is used.

† These σ's were calculated by formula $\left(s = \sqrt{\frac{\Sigma x^2}{N-1}}\right)$ since the samples are small.

Note that the sum of the xy column, namely, 64, differs by 9 from the sum of the xy's in the example above, in which inches and pounds were the units of measurement. However, when deviations are expressed as standard scores, the sum of their products $\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}\right)$ divided by N equals .36, as before.

The quotient

$$\frac{\Sigma\left(\frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y}\right)}{N}$$

is a measure of relationship which remains constant for a given set of data no matter in what units X and Y are expressed. When this ratio is written $\frac{\Sigma xy}{N\sigma_x\sigma_y}$ it becomes the well-known expression for r, the product-moment coefficient of correlation.*

2. The Scatter Diagram and the Correlation Table

When N is small, the ratio method described in the preceding section is often employed for computing the coefficient of correlation between two sets of data. When N is large, however, much time and labor may be saved by first arranging the data in the form of a diagram or chart, and then calculating deviations from assumed, instead of from actual, means. Let us consider the diagram in Figure 47. This chart, which is called a "scatter diagram" or "scattergram," represents the paired heights and weights of 120 college students. The construction of a scattergram is a relatively simple matter. Along the left-hand margin from bottom to top are laid off the class-intervals of the height distribution, measurement expressed in inches; and along the top of the diagram from left to right are laid off the class-intervals of the weight distribution, measurement expressed in pounds. Each of the 120 men is represented on the

* The coefficient of correlation, r, is often called the "Pearson r" after Professor Karl Pearson who developed the product-moment method, following the earlier work of Galton and Bravais. See Walker, H. M., *Studies in the History of Statistical Method* (1929), Chapter 5, pp. 96–111.

Weight in Pounds (X-Variable)

Height in Inches (Y-Variable)	100-109	110-119	120-129	130-139	140-149	150-159	160-169	170-179	f_y	M_{wt}
72-73								1	1	174.5
70-71			1	3	3	4	2	3	16	152.0
68-69			4	11	6	3	2	2	28	142.4
66-67		2	9	11	8	2	1		33	135.1
64-65	1	5	7	10	3				26	128.0
62-63	1	2	7	1	2				13	125.3
60-61	1	1		1					3	117.8
f_x	3	10	28	37	22	9	5	6	120	
M_{ht}	62.5	64.1	65.4	66.6	67.0	68.9	68.9	70.2		

Summary

Weight		Mean ht. for given wt. interval		Height		Mean wt. for given ht. interval	
170–179	Range 80 Lbs.	70.2	Range 7.7 In.	72–73	Range 14 In.	174.5	Range 56.7 Lbs.
160–169		68.9		70–71		152.0	
150–159		68.9		68–69		142.4	
140–149		67.0		66–67		135.1	
130–139		66.6		64–65		128.0	
120–129		65.4		62–63		125.3	
110–119		64.1		60–61		117.8	
100–109		62.5					

FIG. 47. A Scattergram and Correlation Table Showing the Paired Heights and Weights of 120 Students.

diagram with respect to height and weight. Suppose that a man weighs 150 pounds and is 69 inches tall. His weight locates him in the sixth column from the left, and his height in the third row from the top. Accordingly, a "tally" is placed in the third cell of the sixth column. There are three tallies in all in this cell, that is, there are three men who weigh from 150 to 159 pounds, and are 68–69 inches tall. Each of the 120 men

is represented by a tally in a cell or square of the table in accordance with the two characteristics, height and weight. Along the bottom of the diagram in the f_x row is tabulated the number of men who fall in each weight-interval; while along the right-hand margin in the f_y column is tabulated the number of men who fall in each height-interval. The f_y column and f_x row must each total 120, the number of men in all. After all of the tallies have been listed, the frequency in each cell is added and entered on the diagram. The scattergram is then a *correlation table.*

Several interesting facts may be gleaned from the correlation table as it stands. For example, all of the men of a given weight-interval may be studied with respect to the distribution of their heights. In the third column there are twenty-eight men all of whom weigh 120–129 pounds. One of the twenty-eight is 70–71 inches tall; four are 68–69 inches tall; nine are 66–67 inches tall; seven are 64–65 inches tall; and seven are 62–63 inches tall. In the same way, we may classify all of the men of a given height-interval with respect to weight distribution. Thus, in the row next to the bottom, there are thirteen men all of whom are 62–63 inches tall. Of this group one weighs 100–109 pounds; two weigh 110–119 pounds; seven weigh 120–129 pounds; one weighs 130–139 pounds; and two weigh 140–149 pounds. It is fairly clear that the "drift" of paired heights and weights is from the upper right-hand section of the diagram to the lower left-hand section. Even a superficial examination of the diagram reveals a fairly marked tendency for heavy, medium, and light men to be tall, medium, and short, respectively; and this general relationship holds in spite of the scatter of heights and weights within any given "array" (an array is the distribution of cases within a given column or row). Even before making any calculations, then, we should probably be willing to estimate the correlation between height and weight to be positive and fairly high.

Let us now go a step further and calculate the mean height of the three men who weigh 100–109 pounds, the men in column

1. The mean height of this group (using the assumed mean method described in Chapter II, p. 41) is 62.5 inches, and this figure has been written in at the bottom of the correlation table. In the same way, the mean heights of the men who fall in each of the succeeding weight-intervals have been written in at the bottom of the diagram. These data have been tabulated in a somewhat more convenient form below the diagram. From this summary, it appears that an actual weight increase of approximately eighty pounds (180–100) corresponds to an increase in mean height of 7.7 inches; that is, the increase from the lightest to the heaviest man is paralleled by an increase of approximately eight inches in height. It seems clear, therefore, that the correlation between height and weight is positive.

Let us now shift from height to weight, and applying the method used above, find the change in *mean weight* which corresponds to the given change in height.* The mean weight of the three men in the bottom row of the diagram is 117.8 pounds. The mean weight of the thirteen men in the next row from the bottom (who are 62–63 inches tall) is 125.3 pounds. The mean weights of the men who fall in the other rows have been written in their appropriate places in the M_{wt} column. In the summary of results we find that in this group of 120 men an increase of about fourteen inches in height is accompanied by an increase of about 56.7 pounds in mean weight. Thus it appears that the taller the man the heavier he tends to be, and again the correlation between height and weight is seen to be positive.

3. The Graphic Representation of the Correlation Coefficient

It is often helpful in understanding how the correlation coefficient measures relationship to see how a correlation of .00 or .50, say, looks graphically. Figure 48 (1) pictures a correlation of .50. The data in the table are artificial, and were selected to bring out the relationship in as unequivocal a fashion as possible. The scores laid off along the top of the correlation

* This change corresponds to the *second* regression line in the correlation diagram (see p. 280).

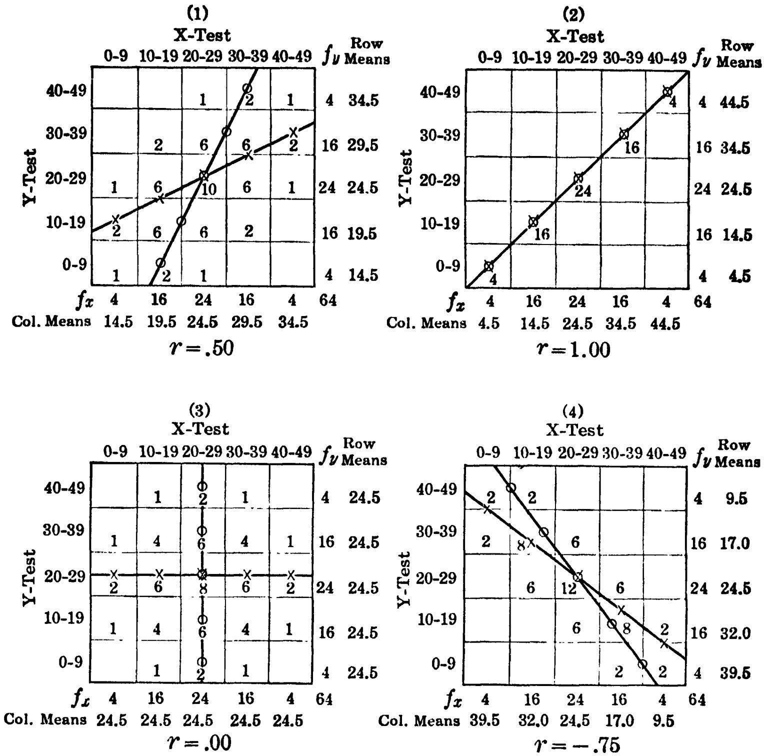


FIG. 48. The Graphical Representation of the Correlation Coefficient.

table from left to right will be referred to simply as the X-test "scores," and the scores laid off at the left of the table from bottom to top as the Y-test "scores." As was done in Figure 47, the mean of each Y-row is entered on the chart, and the means of the X-columns are entered at the bottom of the diagram.

The means of each Y-*array*, that is, the means of the "scores" falling in each X-column, are indicated on the chart by small crosses. Through these crosses a line, called a *regression* line,* has been drawn. This line represents the change in the *mean*

* Regression lines have important properties; they will be defined and discussed more fully in Chapter X.

value of Y over the given range of X. In similar fashion, the means of each X-array, i.e., the means of the scores in each Y-row, are designated on the chart by small circles, through which another line has been drawn. This second *regression* line shows the change in the *mean* value of X over the given range of Y. These two lines together represent the "linear" or straight-line relationship between the variables X and Y.

The closeness of association or degree of correspondence between the X- and Y-tests is indicated by the relative positions of these two regression lines. When the correlation is positive and perfect, the two regression lines close up like a pair of scissors to form one line. Chart (2) in Figure 48 shows how the two regression lines look when $r = 1.00$, and the correlation is perfect. Note that the entries in Chart (2) are concentrated along the diagonal from the upper right- to the lower left-hand section of the diagram. There is no "scatter" of scores in the successive columns or rows, all of the scores in a given array being concentrated within one cell. If Chart (2) represented a correlation table of height and weight, we should know that the tallest man was the heaviest, the next tallest man the next heaviest, and that throughout the group the correspondence of height and weight was perfect.

A very different picture from that of perfect correlation is presented in Chart (3) where the correlation is .00. Here the two regression lines, through the means of the columns and rows, have spread out until they are perpendicular to each other. There is no change in the *mean* Y-score over the whole range of X, and no change in the *mean* X-score over the whole range of Y. This is analogous to the situation described on page 269, in which the mean tapping rate of a group of students was the same for those with "high," "middle," and "low" Army Alpha scores. When the correlation is zero, there is no way of telling from a subject's performance in one test what his performance will be in the other test. The best one can do is to select the *mean* as the most probable value of the unknown score.

Chart (4) in Figure 48 represents a correlation coefficient of

− .75. Negative relationship is shown by the fact that the regression lines, through the means of the columns and rows, run from the upper left- to the lower right-hand section of the diagram. The regression lines are closer together than in Chart (1) where the correlation is .50, but are still separated. If this chart represented a correlation table of height and weight, we should know that the tendency was strong for tall men to be light, and for short men to be heavy.

The charts in Figure 48 represent, as was stated above, a linear relationship between sets of artificial test scores. The data were selected so as to be symmetrical around the means of each column and row, and hence the regression lines go through *all* of the crosses and through *all* of the circles in the successive columns and rows. It is rarely if ever true, however, that the regression lines pass through all of the means of the columns and rows in a correlation table which represents actual test scores or other real measures. Figure 49, which reproduces the correlation table of heights and weights given on page 276, illustrates this fact. The mean heights of the men in the weight (X) columns are indicated by crosses, and the mean weights of the men in the height (Y) rows by circles, as in Figure 48. Note that the series of short lines joining the successive crosses or circles present a decidedly jagged appearance. Two straight lines have been drawn in to describe the general trend of these irregular lines. These two lines go through, or as close as possible to, the crosses or the circles, more consideration being given to those points near the middle of the chart (because they are based upon more data) than to those at the extremes (which are based upon few scores). Regression lines are called lines of "best fit" because they satisfy certain mathematical criteria to be given later (p. 311). Such lines describe better than any other *straight* lines the "run" or "drift" of the crosses and circles across the chart.

In Chapter X we shall develop equations for the "best fitting" lines and show how they may be drawn in to describe the trend of irregular points on a correlation table. For the

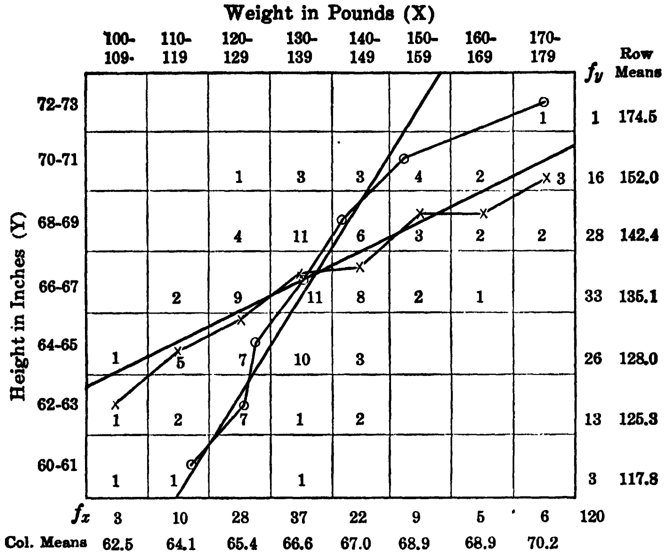


FIG. 49. Graphical Representation of the Correlation between Height and Weight in a Group of 120 College Students. (Fig. 47.)

present, the important fact to get clearly in mind is that when correlation is "linear," the means of the columns and rows in a correlation table can be adequately described by two straight lines and the closer together these two lines the higher the correlation.

III. THE CALCULATION OF THE COEFFICIENT OF CORRELATION BY THE PRODUCT-MOMENT METHOD

1. The Calculation of *r* from a Correlation Table

Having discussed the meaning of correlation in the last sections, we shall now proceed to the calculation of the coefficient of correlation by the product-moment method. Figure 50 will serve as an illustration of the computations required. This correlation table gives the paired heights and weights of 120 college students, and is derived from the scattergram for the same

Height in Inches (Y-Variable)	Weight in Pounds (X-Variable) 100-109	110-119	120-129	130-139	140-149	150-159	160-169	170-179	f_y	y'	fy'	fy'^2	$\Sigma x'y'$ +	$\Sigma x'y'$ −	$\Sigma x'$	$\Sigma x'y'$
72-73								(12) 1 / 12	1	3	3	9	12		4	12
70-71			(-2) 1 / -2	0 3 / 0	(2) 3 / 6	(4) 4 / 16	(6) 2 / 12	(8) 3 / 24	16	2	32	64	58	2	28	56
68-69			(-1) 4 / -4	0 11 / 0	(1) 6 / 6	(2) 3 / 6	(3) 2 / 6	(4) 2 / 8	28	1	28(63)	28	26	4	22	22
66-67		0 2 / 0	0 9 / 0	0 11 / 0	0 8 / 0	0 2 / 0	0 1 / 0		33	0					2	0
64-65	(3) 1 / 3	(2) 5 / 10	(1) 7 / 7	0 10 / 0	(-1) 3 / -3				26	−1	−26	26	20	3	−17	17
62-63	(6) 1 / 6	(4) 2 / 8	(2) 7 / 14	0 1 / 0	(-2) 2 / -4				13	−2	−26	52	28	4	−12	24
60-61	(9) 1 / 9	(6) 1 / 6		0 1 / 0					3	−3	−9(−61)	27	15		−5	15
f_x	3	10	28	37	22	9	5	6	120		2	206	159	−13 (146)	22	146
x'	−3	−2	−1	0	1	2	3	4								
fx'	−9	−20	−28(−57)		22	18	15	24(79)	= 22							
fx'^2	27	40	28		22	36	45	96	= 294							
$\Sigma y'$	−6	−12	−15	2	5	11	6	11	= 2							
$\Sigma x'y'$	18	24	15	0	5	22	18	44	= 146							

check check check

$$c_y = \frac{2}{120} = .02 \qquad c_x = \frac{22}{120} = .18$$

$$c_y^2 = .0004 \qquad c_x^2 = .0324$$

$$\sigma_y = \sqrt{\frac{206}{120} - .0004} \times 2 = 1.31 \times 2 = 2.62$$

$$\sigma_x = \sqrt{\frac{294}{120} - .0324} \times 10 = 1.554 \times 10 = 15.54$$

$$r = \frac{\frac{146}{120} - .02 \times .18}{1.31 \times 1.55}$$

$$r = .60$$

$$PE_r = \frac{.6745\,(1 - .60^2)}{\sqrt{120}} = .04$$

FIG. 50. Calculation of the Product-Moment Coefficient of Correlation between the Heights and Weights of 120 College Students.

data shown in Figure 47. The following outline of the steps in the process of calculating r will be best understood if the student will constantly refer to Figure 50 as he reads through each step.

Step 1

Construct a scattergram for the two variables to be correlated, and from it draw up a correlation table as described on page 275.

Step 2

The distribution of heights for the 120 men is in the f_y column at the right of the diagram. Assume a mean for the height distribution, using the rules given in Chapter II, page 41, and draw double lines to mark off the *row* in which the assumed mean (*ht*) falls. The mean for the height distribution has been taken at 66.5 in. (midpoint of interval 66–67) and the y's have been taken from this point. The prime (′) of the x''s and y''s indicates that these deviations are taken from the *assumed* means of the X and Y distributions (see page 42). Now fill in the fy' and fy'^2 columns. From the first column c_y, the correction in units of interval, is obtained; and this correction together with the sum of the fy'^2 will give the σ of the height distribution, σ_y. As shown by the calculations in Figure 50, the value of σ_y is 2.62 inches.

The distribution of the weights of the 120 men is in the f_x row at the bottom of the diagram. Assume a mean for the weight distribution, and draw double lines to designate the *column* under the assumed mean (*wt*). The mean for the weight distribution is taken at 134.5 pounds (midpoint of interval 130–139), and the x''s are taken from this point. Fill in the fx' and the fx'^2 rows; from the first calculate c_x, the correction in units of interval, and from the second calculate σ_x, the σ of the entire weight distribution. In Figure 50, the value of σ_x is found to be 15.54 pounds.

Step 3

The calculations in *Step 2* simply repeat the now familiar process of calculating σ by the Assumed Mean method. Our first *new* task is to fill in the $\Sigma x'y'$ column at the right of the chart. Since the entries in this column may be either + or −, two columns are provided under $\Sigma x'y'$. Calculation of the entries in the $\Sigma x'y'$ column may be illustrated by considering, first, the single entry in the only occupied cell in the topmost row. The deviation of this cell from the *AM* of the weight distribution, that is, its x', is four *intervals*, and its deviation from the *AM* of the height distribution, that is, its y', is three *intervals*. Hence, the product of the deviations of this cell from the two *AM*'s is 4×3 or 12; and a small figure (12) is placed in the upper right-hand corner of the cell.* The "product-deviation" of the *one* entry in this cell is 1 (4×3) or 12 also, and hence a figure 12 is placed in the lower left-hand corner of the cell. This figure shows the product of the deviations of this single entry from the *AM*'s of the two distributions. Since there are no other entries in the cells of this row, 12 is placed at once under the + sign in the $\Sigma x'y'$ column.

Consider now the next row from the top, taking the cells in order from right to left. The cell immediately below the one for which we have just found the product-deviation also deviates four intervals from the *AM* (*wt*) (its x' is 4), but its deviation from the *AM* (*ht*) is only two intervals (its y' is 2). The product-deviation of this cell, therefore, is 4×2 or 8, as shown by the small figure (8) in the upper right-hand corner of the cell. There are three entries in this cell, and since each has a product-deviation of 8, the final entry in the lower left-hand corner of the cell is $3(4 \times 2)$ or 24. The product-deviation of the second cell in this row is 6 (its x' is 3 and its y' is 2) and since there are two entries in the cell, the final entry is $2(3 \times 2)$ or 12. Each of the

* We may consider the coördinates of this cell to be $x' = 4$, $y' = 3$. The x' is obtained by counting over four intervals from the *vertical column* containing the *AM* (*wt*), and the y' by counting up three intervals from the *horizontal row* containing the *AM* (*ht*). The unit of measurement is the class-interval.

four entries in the third cell over has a product-deviation of 4 (since $x' = 2$ and $y' = 2$) and the final entry is 16. In the fourth cell, each of the three entries has a product-deviation of $2(x' = 1$ and $y' = 2)$ and the cell entry is 6 The entry in the fifth cell over, the cell in the AM (wt) column, is 0, since x' is 0, and accordingly $3(2 \times 0)$ must be 0. Note carefully the entry (-2) in the last cell of the row. Since the deviations of this cell are $x' = -1$, and $y' = 2$, the product $1(-1 \times 2) = -2$, and the final entry is negative. Now we may total up the plus and minus entries in this row and enter the results, 58 and -2, in the $\Sigma x'y'$ column under the appropriate signs.

The final entries in the cells for the other rows of the table and the sums of the product-deviations of each row are obtained as illustrated for the two rows above. The reader should bear in mind in calculating $x'y'$'s that the product-deviations of *all* entries in the cells in the *first* and *third* quadrants of the table are positive, while the product-deviations of *all* entries in the *second* and *fourth* quadrants are negative (p. 11). It should be remembered, too, that all entries either in the column headed by the AM_X or the row headed by the AM_Y have zero product-deviations, since in the one case the x' and in the other the y' equals zero.

Since all entries in a given row have the same y', the arithmetic of calculating $x'y'$'s may often be considerably reduced if each entry in a row-cell is first multiplied by its x', and the sum of these deviations ($\Sigma x'$) multiplied once for all by the common y', viz., the y' of the row. The last two columns $\Sigma x'$ and $\Sigma x'y'$ contain the entries for the rows. To illustrate the method of calculation, in the second row from the bottom, taking the cells in order from right to left, and multiplying the entry in each cell by its x', we have $(2 \times 1) + (1 \times 0) + (7 \times -1) + (2 \times -2) + (1 \times -3)$ or -12. If we multiply this "deviation-sum" by the y' of the whole row (i.e., by -2) the result is 24 which is the final entry in the $\Sigma x'y'$ column. Note that this entry checks the 28 and -4 entered separately in the $\Sigma x'y'$ column by the longer method. This shorter method is often

employed in printed correlation charts and is recommended for use as soon as the student understands fully how the cell entries are obtained.

Step 4 (Checks)

The $\Sigma x'y'$ may be checked by computing the product-deviations and summing for columns instead of rows. The two rows at the bottom of the diagram, $\Sigma y'$ and $\Sigma x'y'$, show how this is done. We may illustrate with the first column on the left, taking the cells from top to bottom. Multiplying the entry in each cell by its appropriate y', we have $(1 \times -1) + (1 \times -2) + (1 \times -3)$ or -6. When this entry in the $\Sigma y'$ row is multiplied by the common x' of the column (i.e., by -3) the final entry in the $\Sigma x'y'$ row is 18. The sum of the $x'y'$ computed from the rows should check the sum of the $x'y'$ computed from the columns.

Two other useful checks are shown in Figure 50. The fy' will equal the $\Sigma y'$ and the fx' will equal the $\Sigma x'$ if no error has been made. The fy' and the fx' are the same as the $\Sigma y'$ and $\Sigma x'$; although these columns and rows are designated differently, they denote in each case the sum of deviations around their AM.

Step 5

When all of the entries in the $\Sigma x'y'$ column have been made, and the column totaled, the coefficient of correlation may be calculated by the formula

$$r = \frac{\dfrac{\Sigma x'y'}{N} - c_x c_y}{\sigma_x \sigma_y} \tag{44}$$

(coefficient of correlation when deviations are taken from the assumed means of the two distributions) *

Substituting 146 for $x'y'$; .02 for c_y; .18 for c_x; 1.31 for σ_y; 1.55 for σ_x; and 120 for N, r is found to be .60. (See Fig. 50.)

* This formula for r differs slightly from the ratio formula developed on page 275. The fact that deviations are taken from assumed rather than from actual means makes it necessary to correct $\Sigma x'y'$ by subtracting the product of the two corrections c_x and c_y.

It is very important to remember that c_x, c_y, σ_x, and σ_y are all left in *units of class-interval* in formula (44). This is done because all product-deviations ($x'y'$'s) are in interval-units, and it is desirable therefore to keep *all* of the terms in the formula in interval-units. Leaving the corrections and the two σ's in units of class-interval facilitates computation, and does not change the result (i.e., the value of the coefficient of correlation).

Several printed charts are available for use in calculating coefficients of correlation by the product-moment method. The following may be mentioned:

1. *The C-D Machine Correlation Chart,* by E. E. Cureton and J. W. Dunlap, published by the Macmillan Co., New York, N.Y.
2. *Dvorak Correlation Chart,* by August Dvorak, published by Longmans, Green and Co., New York, N.Y.
3. *Otis Correlation Chart,* by Arthur Otis, published by the World Book Co., Yonkers, N.Y.
4. *Correlation Chart,* by E. F. Lindquist, published by Houghton Mifflin Co., Boston, Mass.
5. *The Durost-Walker Correlation Chart,* by W. N. Durost and H. M. Walker, published by the World Book Co., Yonkers, N.Y.

2. The Calculation of r from Ungrouped Data

(1) The Formula for r When Deviations Are Taken from the Means of the Two Distributions X and Y

In formula (44) x' and y' deviations are taken from assumed means; and hence it is necessary to correct $\frac{\Sigma x'y'}{N}$ by the product of the two corrections, c_x and c_y (p. 44). When deviations have been taken from the actual means of the two distributions, instead of from assumed means, no correction is needed, as both c_x and c_y are zero. Under these conditions, formula (44) becomes

$$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} \tag{45}$$

(*coefficient of correlation when deviations are taken from the means of the two distributions*)

which is the ratio for measuring correlation developed on page 275. If we write $\sqrt{\frac{\Sigma x^2}{N}}$ for σ_x and $\sqrt{\frac{\Sigma y^2}{N}}$ for σ_y, the N's cancel and formula (45) becomes

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} \tag{46}$$

(*coefficient of correlation when deviations are taken from the means of the two distributions*)

in which x and y are deviations from the actual means as in (45) and Σx^2 and Σy^2 are the sums of the squared deviations in x and y taken from the two means.

When N is fairly large, so that the data can be grouped into a correlation table, formula (44) is always used in preference to formulas (45) or (46) as it entails much less calculation. Formulas (45) and (46) may be used to good advantage, however, in finding the correlation between short, ungrouped series (say, twenty-five cases or so). It is not necessary to tabulate the scores into a frequency distribution. An illustration of the use of formula (46) is given in Table 44. The problem is to find the correlation between the scores made by twelve adults on two tests of "controlled association."

The steps in computing r may be outlined as follows:

Step 1

Find the mean of Test 1 (X) and the mean of Test 2 (Y). The means in Table 44 are 62.5 and 30.4, respectively.

Step 2

Find the deviation of each score on Test 1 from its mean, 62.5, and enter it in column x. Next find the deviation of each score in Test 2 from its mean, 30.4, and enter it in column y.

Step 3

Square all of the x's and all of the y's and enter these squares in columns x^2 and y^2, respectively. Total these columns to obtain Σx^2 and Σy^2.

TABLE 44

TO ILLUSTRATE THE CALCULATION OF r FROM UNGROUPED SCORES WHEN DEVIATIONS ARE TAKEN FROM THE MEANS OF THE SERIES

Subject	Test 1 X	Test 2 Y	x	y	x^2	y^2	xy
A	50	22	− 12.5	− 8.4	156.25	70.56	105.00
B	54	25	− 8.5	− 5.4	72.25	29.16	45.90
C	56	34	− 6.5	3.6	42.25	12.96	− 23.40
D	59	28	− 3.5	− 2.4	12.25	5.76	8.40
E	60	26	− 2.5	− 4.4	6.25	19.36	11.00
F	62	30	− .5	− .4	.25	.16	.20
G	61	32	− 1.5	1.6	2.25	2.56	− 2.40
H	65	30	2.5	− .4	6.25	.16	− 1.00
I	67	28	4.5	− 2.4	20.25	5.76	− 10.80
J	71	34	8.5	3.6	72.25	12.96	30.60
K	71	36	8.5	5.6	72.25	31.36	47.60
L	74	40	11.5	9.6	132.25	92.16	110.40
	750	365			595.00 (Σx^2)	282.92 (Σy^2)	321.50 (Σxy)

$M_X = 62.5 \qquad M_Y = 30.4$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{321.50}{\sqrt{595 \times 282.92}} = .78 \tag{46}$$

Step 4

Multiply the x's and y's in the same rows, and enter these products (with due regard for sign) in the xy column. Total the xy column, taking account of sign, to get Σxy.

Step 5

Substitute for Σxy, 321.50; for Σx^2, 595; and for Σy^2, 282.92 in formula (46), as shown in Table 44, and solve for r.

While formula (46) is useful in calculating r directly from two ungrouped series of scores, it has the same disadvantage as the "long method" of calculating means and σ's described in Chapters II and III. The deviations x and y when taken from the actual means are usually decimals and the multiplication and squaring of these values is often a tedious task. For this reason — even when working with short ungrouped series — it is often easier to assume means, calculate deviations from these AM's, and apply formula (44). The procedure is illustrated in

TABLE 45

TO ILLUSTRATE THE CALCULATION OF r FROM UNGROUPED SCORES WHEN DEVIATIONS ARE TAKEN FROM THE ASSUMED MEANS OF THE SERIES

Subject	Test 1 X	Test 2 Y	x'	y'	x'^2	y'^2	$x'y'$
A	50	22	− 10	− 8	100	64	80
B	54	25	− 6	− 5	36	25	30
C	56	34	− 4	4	16	16	− 16
D	59	28	− 1	− 2	1	4	2
E	60	26	0	− 4	0	16	0
F	62	30	2	0	4	0	0
G	61	32	1	2	1	4	2
H	65	30	5	0	25	0	0
I	67	28	7	− 2	49	4	− 14
J	71	34	11	4	121	16	44
K	71	36	11	6	121	36	66
L	74	40	14	10	196	100	140
	750	365			670 ($\Sigma x'^2$)	285 ($\Sigma y'^2$)	334 ($\Sigma x'y'$)

$AM_X = 60.0$ $\quad AM_Y = 30.0$

$M_X = 62.5$ $\quad M_Y = 30.4$

$c_x = 2.5$ $\quad c_y = .4$

$c^2_x = 6.25$ $\quad c^2_y = .16$

$$r = \frac{\frac{334}{12} - 1.00}{7.04 \times 4.86} \tag{44}$$

$\sigma_x = \sqrt{\frac{670}{12} - 6.25}$ $\quad \sigma_y = \sqrt{\frac{285}{12} - .16}$

$= 7.04$ $\quad = 4.86$ $\quad r = .78$

Table 45 with the same data given in Table 44. Note that the two means, M_X and M_Y, are first calculated. The corrections, c_x and c_y, are found by subtracting AM_X from M_X and AM_Y from M_Y (p. 44). Since deviations are taken from assumed means, fractions are avoided; and the calculations of $\Sigma x'^2$, $\Sigma y'^2$, $\Sigma x'y'$ are readily made. Substitution in formula (44) then gives r.

(2) The Calculation of r from Raw Scores, i.e., When Deviations Are Taken from Zero

The calculation of r may often be carried out most readily — especially when a calculating machine is available — by means of the following formula which is based upon "raw" or obtained scores:

$$r = \frac{\Sigma XY - NM_XM_Y}{\sqrt{[\Sigma X^2 - NM^2_X][\Sigma Y^2 - NM^2_Y]}} \tag{47}$$

(*coefficient of correlation calculated from raw or obtained scores*)

In this formula, X and Y are obtained scores, and M_X and M_Y are the means of the X and Y series, respectively. ΣX^2 and ΣY^2 are the sums of the squared X and Y values, and N is the number of cases.

Formula (47) is derived directly from formula (44) by assuming the means of the X and Y tests to be zero. If AM_X and AM_Y are zero, each X and Y score is a deviation from its AM as it stands, and hence we work with the scores themselves. Since the correction, c, always equals $M - AM$, it follows that when the AM equals 0, $c_x = M_X$, $c_y = M_Y$ and $c_xc_y = M_XM_Y$. Furthermore, when $c_x = M_X$ and $c_y = M_Y$ and the "scores" are "deviations," the formula

$$\sigma_x = \sqrt{\frac{\Sigma fx'^2}{N} - c^2_x} \times \text{interval}$$

(see p. 62) becomes

$$\sigma_x = \sqrt{\frac{\Sigma X^2}{N} - M^2_X}$$

and σ_y for the same reason equals $\sqrt{\frac{\Sigma Y^2}{N} - M^2_Y}$ If we substitute these values for c_xc_y, σ_x, and σ_y in formula (44), the formula for r in terms of raw scores given in (47) is obtained.

An alternate form of (47) is often more useful in practice. This is

$$r = \frac{N\Sigma XY - \Sigma X \times \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \tag{48}$$

(*coefficient of correlation calculated from raw or obtained scores*)

This formula is obtained from (47) by substituting $\frac{\Sigma X}{N}$ for M_X, and $\frac{\Sigma Y}{N}$ for M_Y in numerator and denominator, and canceling the N's.

The calculation of r from original scores is shown in Table 46.

The data are again the two sets of twelve scores obtained on the "controlled association" tests, the correlation for which was found to be .78 in Table 44. This short example is for the pur-

TABLE 46

To Illustrate the Calculation of r from Ungrouped Data When Deviations Are Original Scores (AM's = 0)

Subject	Test 1 X	Test 2 Y	X^2	Y^2	XY
A	50	22	2500	484	1100
B	54	25	2916	625	1350
C	56	34	3136	1156	1904
D	59	28	3481	784	1652
E	60	26	3600	676	1560
F	62	30	3844	900	1860
G	61	32	3721	1024	1952
H	65	30	4225	900	1950
I	67	28	4489	784	1876
J	71	34	5041	1156	2414
K	71	36	5041	1296	2556
L	74	40	5476	1600	2960
	750	365	47470	11385	23134

$M_X = 62.50$
$M_Y = 30.42$ (means to two decimals)

$$r = \frac{23134 - 12 \times 62.50 \times 30.42}{\sqrt{[47470 - 12 \times (62.50)^2][11385 - 12 \times (30.42)^2]}} \tag{47}$$

$r = .78$

pose of illustrating the arithmetic and must not be taken as a recommendation that formula (47) be used only with short series. As a matter of fact, formula (47) or (48) is most useful, perhaps, with long series, especially if one is working with a calculating machine.

The computation by formula (48) is straightforward and the method easy to follow, but the calculations become tedious if the scores are expressed in more than two digits. For this reason, when using formula (48) it will often greatly lessen the arithmetical work, if we first "reduce" the original scores by subtracting a constant quantity from each of the original X and Y scores. In Table 47, the same two series of twelve scores have been reduced by subtracting 65 from each of the X scores, and 25 from each of the Y scores. The reduced scores, entered in

TABLE 47

TO ILLUSTRATE THE CALCULATION OF r FROM UNGROUPED DATA WHEN DEVIATIONS ARE ORIGINAL SCORES (AM's = 0)

Scores are "reduced" by the subtraction of 65 from each X, and 25 from each Y to give X' and Y'.

Subject	Test 1 X	Test 2 Y	X'	Y'	X'^2	Y'^2	$X'Y'$
A	50	22	− 15	− 3	225	9	45
B	54	25	− 11	0	121	0	0
C	56	34	− 9	9	81	81	− 81
D	59	28	− 6	3	36	9	− 18
E	60	26	− 5	1	25	1	− 5
F	62	30	− 3	5	9	25	− 15
G	61	32	− 4	7	16	49	− 28
H	65	30	0	5	0	25	0
I	67	28	2	3	4	9	6
J	71	34	6	9	36	81	54
K	71	36	6	11	36	121	66
L	74	40	9	15	81	225	135
	750	365	− 30($\Sigma X'$)	65($\Sigma Y'$)	670($\Sigma X'^2$)	635($\Sigma Y'^2$)	159($\Sigma X'Y'$)

$$M_X = \frac{\Sigma X'}{N} + 65 = -\frac{30}{12} + 65 = 62.5 \qquad M_Y = \frac{\Sigma Y'}{N} + 25 = \frac{65}{12} + 25 = 30.$$

$$r = \frac{(12 \times 159) + (30 \times 65)}{\sqrt{[12 \times 670 - (-30)^2][12 \times 635 - (65)^2]}} \tag{48}$$

$$= \frac{3858}{4923} = .78$$

the table under X' and Y', are first squared to give $\Sigma X'^2$ and $\Sigma Y'^2$, and then multiplied by rows to give $\Sigma X'Y'$. Substitution of these values in formula (48) gives the coefficient of correlation r. If the means of the two series are wanted, these may readily be found by adding to $\frac{\Sigma X'}{N}$ and $\frac{\Sigma Y'}{N}$ the amounts by which the X and Y scores were reduced (see computations in Table 47).

The method of computing r by first reducing the scores is

usually superior to the method of applying formula (47) or (48) directly to the raw scores. This is because we deal with smaller whole numbers, and much of the arithmetic can be done mentally. When raw scores have more than two digits, they are cumbersome to square and multiply unless reduced. The student should note that instead of 65 and 25 other constants might have been used to reduce the X and Y scores. If the smallest X and Y scores had been subtracted, namely, 50 and 22, all of the X' and Y' would, of course, have been positive. This is an advantage in calculation but these reduced scores would have been somewhat larger numerically than are the reduced scores in Table 47. In general, the best plan in reducing scores is to subtract constants which are close to the means. The reduced scores are then both plus and minus, but are numerically about as small as we can make them.

(3) The Calculation of r by the Difference-Formula

It is apparent from the preceding sections that the product-moment formula for r may be written in several ways, depending upon whether deviations are taken from actual or assumed means, and upon whether raw scores or deviations are employed. The present section contributes still another formula for calculating r — namely, the difference-formula. This formula will complete our list of expressions for r, as it is believed that the student who understands the meaning and use of the correlation formulas given in this chapter will have no difficulty with other variations which he may encounter.*

The formula for r by the difference method is

$$r = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma d^2}{2\sqrt{\Sigma x^2 \times \Sigma y^2}} \tag{49}$$

(coefficient of correlation by difference-formula, deviations from the means of the distributions)

in which $\Sigma d^2 = \Sigma(x - y)^2$.

* See the following article which lists fifty-two variations of the r-formula: Symonds, P. M., "Variations of the Product-Moment (Pearson) Coefficient of Correlation," *Journal of Educ. Psych.* 17 (1926), 458–469.

The principal advantage of the difference-formula is that no cross products (xy's) need be computed. For this reason, this formula is employed in several of the printed correlation charts. Formula (49) is illustrated in Table 48 with the same data used in Table 44 and elsewhere in this chapter. Note that the x, y, x^2, and y^2 columns repeat Table 44. The d or $(x - y)$ column is found by subtracting algebraically each y-deviation from its corresponding x-deviation. These differences are then squared and entered in the d^2 or $(x - y)^2$ column. Substitution of Σx^2, Σy^2, and Σd^2 in formula (49) gives $r = .78$.

TABLE 48

To Illustrate the Calculation of r from Ungrouped Data by the Difference-Formula, Deviations from the Means

Subject	Test 1 X	Test 2 Y	x	y	d $(x - y)$	x^2	y^2	d^2 $(x - y)^2$
A	50	22	− 12.5	− 8.4	− 4.1	156.25	70.56	16.81
B	54	25	− 8.5	− 5.4	− 3.1	72.25	29.16	9.61
C	56	34	− 6.5	3.6	− 10.1	42.25	12.96	102.01
D	59	28	− 3.5	− 2.4	− 1.1	12.25	5.76	1.21
E	60	26	− 2.5	− 4.4	1.9	6.25	19.36	3.61
F	62	30	− .5	− .4	− .1	.25	.16	.01
G	61	32	− 1.5	1.6	− 3.1	2.25	2.56	9.61
H	65	30	2.5	− .4	2.9	6.25	.16	8.41
I	67	28	4.5	− 2.4	6.9	20.25	5.76	47.61
J	71	34	8.5	3.6	4.9	72.25	12.96	24.01
K	71	36	8.5	5.6	2.9	72.25	31.36	8.41
L	74	40	11.5	9.6	1.9	132.25	92.16	3.61
						595.00	282.92	234.92

$M_X = 62.5$

$M_Y = 30.4$

$$r = \frac{595.00 + 282.92 - 234.92}{2\sqrt{595 \times 282.92}} \qquad (49)$$

$$= .78$$

Another form of the difference-formula is often useful, especially in machine calculation. This version makes use of raw or obtained scores:

$$r = \frac{N[\Sigma X^2 + \Sigma Y^2 - \Sigma(X - Y)^2] - 2(\Sigma X) \times (\Sigma Y)}{2\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \qquad (50)$$

(coefficient of correlation by difference-formula, calculation from raw or obtained scores)

in which $\Sigma(X - Y)^2$ is the sum of the squared differences between the two sets of scores.

IV. Reliability of the Coefficient of Correlation

1. The Standard and Probable Errors of a Coefficient of Correlation

The usual formulas for the standard and probable errors of a coefficient of correlation are

$$\sigma_r = \frac{(1 - r^2)}{\sqrt{N - 1}} \tag{51}$$

and

$$PE_r = \frac{.6745(1 - r^2)}{\sqrt{N - 1}} \tag{52}$$

(*standard and probable errors of a coefficient of correlation*)

The PE_r formula is the more often used, perhaps because the PE has become established in the literature as the result of long usage. When $r = .60$ and $N = 120$ (see height and weight problem in Fig. 47), $PE = .04$ to two decimals [from (52)]. This probable error is taken to mean that the chances are 50 in 100 (odds 1:1) that the obtained r of .60 does not miss the true or population value by more than $\pm$ .04.

There are two serious objections to the use of formulas (51) and (52). In the first place, the r in these formulas is really the true or population r. Since we do not have the true r, we must substitute the calculated or sample r in the formula in order to get an estimate of the standard or probable error. If the obtained r is in error, our estimate will also be in error; and at best it is approximate.

In the second place, the sampling distribution of r is not normal except when the population r is .00 and N is large. When r is high (.80 or more) and N is small, the sampling distribution of r is skewed and the PE is decidedly misleading. The reason for skewness in the sampling distribution of high r's grows out of the fact that the range of r's is definitely limited at + 1.00

and -1.00. Suppose, for example, that $r = .80$ and $N = 20$. Then in a new sample of twenty cases the probability of an r less than .80 is much greater than the probability of an r greater than .80 because of the obtained r's nearness to unity. The distribution of r's obtained from successive samples of twenty cases will be skewed negatively (p. 119) and the skewness to the left will increase as r increases. For small and intermediate values of r, say between $\pm .50$, and for N's of 100 or more, the distribution of r in successive samples will conform fairly closely to the normal curve and formulas (51) and (52) will yield useful estimates of reliability. But unless used with caution, PE_r is likely to be misleading.

It has been customary for a long time to regard an r as worthy of confidence if it is at least four times its PE. If $r = .20$ and $N = 40$, $PE_r = .10$, and our r is twice its PE. On the assumption that the true r in the population is zero, the obtained r of .20 (since it is only $2PE$ from zero) could well be attributed to sampling errors, and hence is not significant. When N is 150, however, the correlation coefficient of .20 is four times its probable error of .05 and can hardly be attributed solely to accidents of sampling.

2. Testing the Reliability of a Coefficient of Correlation Against the Null Hypothesis

The significance of an obtained r may be tested more exactly against the null hypothesis than in terms of PE_r. Assuming the population r to be zero, the method consists in comparing the t value (see Table 29) for the obtained r with the t's to be expected by chance at the .05 and .01 limits. The t for a given r is found from the formula

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{53}$$

(t for determining the significance of a computed r on the null hypothesis)

in which r = the obtained coefficient and N = the number of cases. The value of t may be read from Table 29, page 190,

which is entered with $N - 2$ degrees of freedom. To illustrate, suppose $r = .60$ and $N = 120$ (p. 283). Then from (53) $t = \frac{.60\sqrt{118}}{.80}$ or 8.15. Entering Table 29 with 118 degrees of freedom ($N - 2 = 118$), we find that t at the .05 level is 1.98, and at the .01 level, 2.62. Since our t is far larger than the second of these values, we conclude forthwith that the null hypothesis is clearly disproved and our r is very significant. The probability that we should have obtained an r of .60, if the true r were .00, is much less than .01.

A simpler method of testing the significance of an r than by computing t is to enter Table 49 with $N - 2$ degrees of freedom

TABLE 49

CORRELATION COEFFICIENTS AT THE 5% AND 1% LEVELS OF SIGNIFICANCE

Example: When N is 52 and $(N - 2)$ is 50, an r must be .273 to be significant at .05 level, and .354 to be significant at .01 level.

Degrees of freedom $(N - 2)$	.05	.01	Degrees of freedom $(N - 2)$	.05	.01
1	.997	1.000	24	.388	.496
2	.950	.990	25	.381	.487
3	.878	.959	26	.374	.478
4	.811	.917	27	.367	.470
5	.754	.874	28	.361	.463
6	.707	.834	29	.355	.456
7	.666	.798	30	.349	.449
8	.632	.765	35	.325	.418
9	.602	.735	40	.304	.393
10	.576	.708	45	.288	.372
11	.553	.684	50	.273	.354
12	.532	.661	60	.250	.325
13	.514	.641	70	.232	.302
14	.497	.623	80	.217	.283
15	.482	.606	90	.205	.267
16	.468	.590	100	.195	.254
17	.456	.575	125	.174	.228
18	.444	.561	150	.159	.208
19	.433	.549	200	.138	.181
20	.423	.537	300	.113	148
21	.413	.526	400	.098	.128
22	.404	.515	500	.088	.115
23	.396	.505	1000	.062	.081

and compare our sample r with the tabulated entries. Two significance levels, .05 and .01, appear in Table 49. The table is read as follows: Suppose $r = .60$ and $N = 120$. Then for 118 degrees of freedom the entries at .05 and .01 are by linear interpolation, .180 and .235, respectively. This means that only five times in 100 trials would an r as large as $\pm$ .180 appear by accidents of sampling if the population r were actually .00; and only once in 100 trials would an r of $\pm$ .235 appear if the popula-

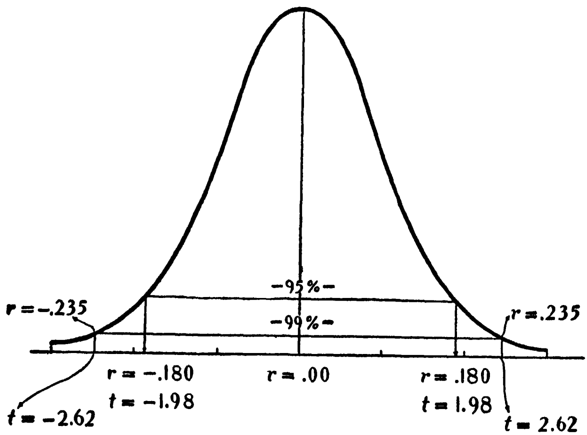


FIG. 51.

When the true r is zero and $N = 120$ (118 *df*) 5% of sample r's exceed $\pm$.180, and 1% exceed $\pm$ .235.

tion r were .00. It is clear that the obtained r of .60, since it is much larger than .235, is very significant. Another way of stating the same conclusion is to say that we may be confident at the .01 level that the true r is *not* zero. Figure 51 represents the situation outlined in the example above. The entries in Table 49 were found by substituting for N and for t in (53), the t's being taken from the .05 and .01 columns in Table 29.

It will be noted from Figure 51 that Table 49 takes account of *both* ends of the sampling distribution — does not consider the sign of r. When $N = 120$, the probability ($P/2$) of an r of + .180 or *more*, on the null hypothesis, is .025; and the proba-

bility ($P/2$) of an r of $-.180$ or *less* is also .025. For a $P/2$ of .01 (or $P = .02$), the r by interpolation between .05 (.180) and .01 (.235) is .221. On the null hypothesis, therefore, only once in 100 trials would a *positive* r of .221 or a *larger* value arise through sampling accidents.

The .05 and .01 levels in Table 49 are the only ones we will need ordinarily in evaluating the significance of a calculated r. Several illustrations of the use of Table 49 are given below:

Size of Sample (N)	Degrees of Freedom ($N - 2$)	Calculated r	Interpretation
10	8	.70	significant at .05, not at .01 level
152	150	$-.12$	not significant
27	25	.50	significant at .05, hardly at .01 level
500	498	.20	very significant
100	98	$-.30$	very significant

It is clear from these examples that even a small r may be significant if computed from a large sample, while an r as high as .70 may not be very significant if N is small. Table 49 is especially useful when N is small, as it is here that the PE of an r is most apt to be misleading. Suppose, for example, that we have calculated an r of .55 for a sample of twelve cases. The PE of this r, by formula (52) is .14, and since the r of .55 is about four times its PE, we might conclude that our correlation is very significant. From Table 49, however, we note that for 10 degrees of freedom ($N - 2 = 10$), an r must be .708 to be significant at the .01 level. Furthermore, an r must be .642 before the probability is .01 that this r or a larger value will occur on the null hypothesis (at $P = .02$, $r = .642$ by interpolation between .05 and .01 in Table 49). For this small sample, a conclusion as to significance based upon the PE_r would clearly be in error.

The interpretation of the significance of a low r should always be tentative. Even when small r's are significant by our tests,

it is a good plan to repeat the experiment on another sample before announcing a final decision.

3. Testing the Reliability of a Correlation Coefficient by Fisher's z-Function

R. A. Fisher * has shown that r can be transformed into a new statistic called z which is normally or nearly normally distributed (p. 297) no matter what the size of r. A further advantage of z is that its standard error depends entirely upon the size of the sample and is independent of the calculated value of z. The z corresponding to any given r may be read from a table provided by Fisher.

The significance of any given r may be determined by transforming it into a z, calculating the SE of z, and applying tests of significance. If z divided by SE_z is greater than 2.58 (Table 29) the null hypothesis may be safely discarded. The transformed r or z may also be used in testing the significance of the difference between two r's. When our r's are obtained from *independent* random samples, formula (29) may be used, the SE's of the z's being substituted in the formula for σ_M's. When two or more r's are obtained from the *same* sample, the z transformation is no longer strictly applicable, but an approximate method may still be employed.†

4. Averaging Coefficients of Correlation

It is a fairly common practice to average correlation coefficients computed from tests given to comparable groups in order to obtain a generalized picture of the relationship between the two variables. The averaging of r's is a dubious and often an incorrect procedure. (1) r's do not vary along a linear scale so that the increase from .40 to .50 does not mean the same increase in relationship as does an increase from .80 to .90. (2) When

* Fisher, R. A., *Statistical Methods for Research Workers* (8th ed., 1941), pp. 190–203.

† Lindquist, E. F., *Statistical Analysis in Educational Research* (1940), pp. 217–218.

$+ r$'s and $- r$'s are averaged, they tend to cancel each other out. Thus the mean of an r of .60 and an r of $-$ .60 is .00, and two substantial measures of correlation combine to give a result which indicates no real relationship. When r's do not differ greatly in size, an arithmetic mean will yield a result which is often useful; but this is not true when r's differ widely in size or in sign. Averaging an r of .70 and an r of .60 to obtain .65 is permissible; but averaging an r of .90 and an r of .10 to obtain .50 is not.

The safest plan is not to average r's at all. When for various reasons averaging seems to be demanded by the problem, the best method is to transform the r's into z's and take the arithmetic mean of the z's. An average r may then be obtained from the average z.

PROBLEMS

1. Find the correlation between the two sets of scores given below, using the ratio method (p. 272).

Subjects	X	Y
a	15	40
b	18	42
c	22	50
d	17	45
e	19	43
f	20	46
g	16	41
h	21	41

2. The scores given below were achieved upon Army Alpha and Typewriting Tests by 100 students in a typewriting class. The typewriting scores are in number of words written per minute, with certain penalties. Find the coefficient of correlation and its PE_r. Check the significance of r by Table 49. Use an interval of five units for Y and an interval of ten units for X.

Typing (Y)	Alpha (X)	Typing (Y)	Alpha (X)	Typing (Y)	Alpha (X)
46	152	26	164	40	120
31	96	33	127	36	140
46	171	44	144	43	141
40	172	35	160	48	143
42	138	49	106	45	138
41	154	40	95	58	149
39	127	57	146	23	142
46	156	23	175	45	166
34	156	51	126	44	138
48	133	35	120	47	150
48	173	41	154	29	148
38	134	28	146	46	166
26	179	32	154	46	146
37	159	50	159	39	167
34	167	29	175	49	139
51	136	41	164	34	183
47	153	32	111	41	150
39	145	49	164	49	179
32	134	58	119	31	138
37	184	35	160	47	136
26	154	48	149	40	172
40	90	40	149	30	145
53	143	43	143	40	109
46	173	38	159	38	158
39	168	37	157	29	115
52	187	41	153	43	93
47	166	51	149	55	163
31	172	40	163	37	147
33	189	35	175	52	169
22	147	31	133	38	75
46	150	23	178	39	152
44	150	37	168	32	159
37	143	46	156	42	150
31	133				

3. In the correlation table given below compute the coefficient of correlation. Test the significance of r (p. 297).

BOYS: AGES 4.5 TO 5.5 YEARS

Weight in Pounds (X)

Height in Inches (Y)	24–28	29–33	34–38	39–43	44–48	49–53	Totals
45–47			1		2		3
42–44			4	35	21	5	65
39–41		5	87	90	7	1	190
36–38	1	18	72	8			99
33–35	5	15	5				25
30–32	2						2
Totals	8	38	169	133	30	6	384

4. In the following correlation table compute the coefficient of correlation and test its significance.

Army Alpha I.Q.'s

School Marks	84 and lower	85–89	90–94	95–99	100–104	105–109	110–114	115–119	120–124	125 over	Totals
90 and over				3	3	15	12	9	9	5	56
85–89				8	17	15	24	13	6	6	89
80–84			4	6	22	21	20	10	5	1	89
75–79			7	25	33	23	10	7	4		109
70–74		4	10	18	14	22	12	1	1		82
65–69	1	3	3	12	7	8	8	1			43
60–64			2	5	3	1	1				12
Totals	1	7	26	77	99	105	87	41	25	12	480

5. Compute the coefficient of correlation between the Algebra Test scores and I.Q.'s shown in the table below and test its significance.

ALGEBRA TEST SCORES

I.Q.'s	30–34	35–39	40–44	45–49	50–54	55–59	60–64	65–69	Totals
130–139				1		1		1	3
120–129			1		1	2	1		5
110–119	1	2	5	6	11	6	3	2	36
100–109	3	7	9	17	13	5	1	1	56
90–99	4	10	16	12	5	1			48
80–89	4	9	8	2	2				25
Totals	12	28	39	38	32	15	5	4	173

6. Compute the correlation between the two sets of scores given below
   (*a*) when deviations are taken from the means of the two series [use formula (46)];
   (*b*) when the means are taken at zero. First reduce the scores by subtracting 150 from each of the scores in Test 1, and 40 from each of the scores in Test 2.
   (*c*) Test the significance of r.

Test 1	Test 2	Test 1	Test 2
150	60	139	41
126	40	155	43
135	45	147	37
176	50	162	58
138	56	156	48
142	43	146	39
151	57	133	31
163	38	168	46
137	41	153	52
178	55	150	57

7. Find the correlation between the two sets of memory-span scores given below (the first series is arranged in order of size) (*a*) when deviations are taken from assumed means [formula (44)], (*b*) by the difference-method given on page 295. Test significance of r.

Test 1 (digit span)	Test 2 (letter span)
15	12
14	14
13	10
12	8
11	12
11	9
11	12
10	8
10	10
10	9
9	8
9	7
8	7
7	8
7	6

8. Fill in the following table:

	Size of Sample (N)	Degrees of Freedom ($N - 2$)	r	Significance
(*a*)	15	13	$-.68$	
(*b*)	30	28	.22	
(*c*)	82	80	$-.30$	
(*d*)	225	223	.05	

Answers

1. $r = .60$
2. $r = -.05$; $PE_r = .07$; not significant
3. $r = .71$; highly significant, beyond .01 level
4. $r = .46$; highly significant, beyond .01 level
5. $r = .52$; highly significant, beyond .01 level

6. $r = .41$; r not significant at .05 level
7. $r = .78$; significant beyond .01 level
8. (*a*) very significant (beyond .01 level)
   (*b*) not significant
   (*c*) very significant
   (*d*) not significant

CHAPTER X

REGRESSION AND PREDICTION

I. The Regression Equations

1. The Problem of Predicting One Variable from Another

Suppose that in a group of 120 college students (p. 283), we wish to estimate a certain man's height knowing his weight to be 153 pounds. The best possible "guess" that we can make of this man's height is the mean height of all of the men who fall in the 150–159 weight-interval. In Figure 52 the mean height of the nine men in this column is 68.9 inches, which is, therefore, the most *likely* height of a man who weighs 153 pounds. In the same way, the most probable height of a man who weighs 136 pounds is 66.6 inches, the mean height of the thirty-seven men who fall in weight-column 130–139 pounds. And, in general, the most probable height of any man in the group is the *mean* of the heights of *all* of the men who weigh the same (or approximately the same) as he, i.e., who fall within the same weight-column.

Turning to weight, we can make the same kind of estimates. Thus, the best possible "guess" that we can make of a man's weight knowing his height to be 66.5 inches is 135.1 pounds, viz., the mean weight of the thirty-three men who fall in the height-interval 66–67 inches. Again, in general, the most probable weight of any man in the group is the *mean* weight of *all* of the men who are of the same (or approximately the same) height.

Our illustration shows that from the scatter diagram alone it is possible to "predict" one variable from another. But the prediction is rough, and is obviously subject to a large "error of estimate." * Moreover, while we have made use of the fact

* See page 320.

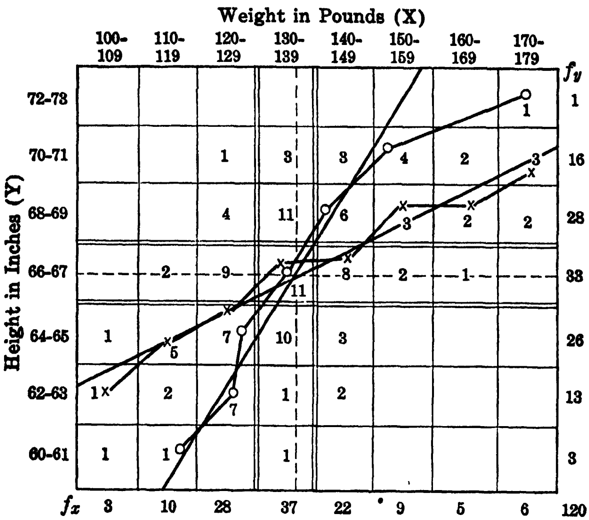


FIG. 52. Illustrating Positions of Regression Lines and Calculation of the Regression Equations. (See Fig. 50, p. 283.)

$r = .60$
$M_X = 136.3$ lbs.
$M_Y = 66.5$ inches

For plotting on the chart, regression equations are written with σ_x and σ_y in class-interval units, viz. —

$$\left.\begin{aligned} y &= .51x \\ x &= .71y \end{aligned}\right\} \text{ see p. 316.}$$

Calculation of Regression Equations

I. Deviation Form

(1) $$\bar{y} = .60 \times \frac{2.62}{15.54} x = .10x \quad (54)$$

(2) $$\bar{x} = .60 \times \frac{15.54}{2.62} y = 3.56y \quad (55)$$

II. Score Form

(1) $Y - 66.5 = .10(X - 136.3)$ or $\overline{Y} = .10X + 52.9$ (56)

(2) $X - 136.3 = 3.56(Y - 66.5)$ or $\overline{X} = 3.56Y - 100.4$ (57)

Calculation of Standard Errors of Estimate

$$\sigma_{(\text{est. } Y)} = 2.62\sqrt{1 - .60^2} = 2.10 \text{ inches} \quad (58)$$

$$\sigma_{(\text{est. } X)} = 15.54\sqrt{1 - .60^2} = 12.43 \text{ pounds} \quad (59)$$

that the means are the most probable points in our arrays (columns or rows), we have made no use of our knowledge concerning the correlation between the two variables. The two regression lines * in Figure 52 are definitely determined by the correlation between height and weight and their degree of separation indicates the size of the correlation coefficient (p. 279). Consequently, they describe more regularly, and in a more generalized fashion than do the series of short straight lines joining the means, the relationship between height and weight *over the whole range* (see also p. 282). A knowledge of the equations of these lines is necessary if we are to make as accurate a prediction as our data will permit. For example, given the weight (X) of a man comparable to those in our group, on substituting in the equation connecting Y and X we are able to predict this man's height more accurately than if we took the mean of his height array. The task of the next section will be to develop equations for the two regression lines by means of which precise predictions from X to Y or from Y to X can be accomplished.

2. The Two Regression Equations in Deviation Form

The equations of the two regression lines in a correlation table represent the straight lines which "best fit" the means of the successive columns and rows in the table. Using as a definition of "best fit" the criterion of "least squares," † Pearson worked out the equation of the line which goes through, or as close as possible to, more of the column-means than any other straight

* The term "regression" was first used by Francis Galton with reference to the inheritance of stature. Galton found that children of tall parents tend to be less tall, and children of short parents less short, than their parents. In other words, the heights of the offspring tend to "move back" toward the mean height of the general population. This tendency toward maintaining the "mean height" Galton called the principle of regression, and the line describing the relationship of height in parent and offspring was called a "regression line." The term is still employed, although its original meaning of "stepping back" to some stationary average is not necessarily implied (see p. 331).

† For an elementary mathematical treatment of the method of least squares as applied to the problem of fitting regression lines, see Walker, H. M., *Elementary Statistical Method* (1943), pp. 308–310.

line; and also the equation of the line which goes through, or as close as possible to, more of the row-means than any other straight line. These two lines are "best fitting" in a mathematical sense, the one to the observations of the columns and the other to the observations of the rows.

The equation of the first regression line, the line drawn to represent the crosses in Figure 52, is as follows:

$$\bar{y} = r\,\frac{\sigma_y}{\sigma_x} \times x \tag{54}$$

(*regression equation of* y *on* x, *deviations taken from the means of* Y *and* X)

The factor $r\,\frac{\sigma_y}{\sigma_x}$ is called the *regression coefficient*, and is often replaced in (54) by the term b_{yx} or b_{12} so that formula (54) may be written $\bar{y} = b_{yx} \times x$, or $\bar{y} = b_{12} \times x$. The bar over the $(\bar{y})$ means that our estimate is an average value.

If we substitute in formula (54) the values of r, σ_y, and σ_x, obtained from Figure 52, we have

$$\bar{y} = .60 \times \frac{2.62}{15.54}\,x, \text{ or } \bar{y} = .10x$$

This equation gives the relationship of deviations from mean height to deviations from mean weight. When $x = 1.00$, $\bar{y} = .10$; and a deviation of one pound from the mean of the X's (weight) is accompanied by a deviation of .10 inch from the mean of the Y's (height). The man who stands one pound *above* the mean weight of the group, therefore, is most probably .10 inch *above* the mean height. Since this man's weight is 137.3 pounds (136.3 + 1.00), his height is most probably 66.6 inches (66.5 + .10). Again, the man who weighs 120 pounds, i.e., is 16.3 pounds *below* the mean of the group, is most probably 64.9 inches tall — or about 1.6 inches *below* the mean height of the group. To get this last value, substitute $x = -16.3$ in the equation above to get $\bar{y} = -1.63$, and refer this value to its mean. The regression equation is, in effect, a generalized statement. It tells us that the most probable deviation of an indi-

vidual in our group from the M (*ht*) is just .10 of his deviation from the M (*wt*).

The equation $\bar{y} = r\,\frac{\sigma_y}{\sigma_x} \times x$ gives the relationship between y and x in *deviation form.* This designation is necessary because the two variables are expressed as deviations from their respective means (i.e., as x and y); hence, for a given *deviation* from M_X the equation gives the most probable accompanying *deviation* from M_Y.

The equation of the second regression line, the line drawn through the means of the rows in Figure 52, is

$$\bar{x} = r\,\frac{\sigma_x}{\sigma_y} \times y \qquad (55)$$

(*regression equation of* x *on* y, *deviations taken from the means of* X *and* Y)

As in the first regression equation, the regression coefficient $r\,\frac{\sigma_x}{\sigma_y}$ is often replaced by the expression b_{xy} or b_{21} and formula (55) written $\bar{x} = b_{xy} \times y$ or $\bar{x} = b_{21} \times y$.

If we substitute for r, σ_x, and σ_y, in formula (55), we have

$$\bar{x} = .60 \times \frac{15.54}{2.62}\,y \text{ or } \bar{x} = 3.56y$$

from which it is evident that a deviation of 1 inch from the M (*ht*), or from 66.5 inches, is accompanied by a deviation of 3.56 pounds from the M (*wt*), or from 136.3 pounds. Expressed generally, the most probable deviation of *any man* from the mean weight is just 3.56 times his deviation from the mean height. Accordingly, a man 67 inches tall or .5 inch *above* the mean height $(66.5 + .5 = 67)$ most probably weighs 138.1 pounds, or is 1.8 pounds *above* the mean weight $(136.3 + 1.8)$. (Substitute $y = .5$ in the equation and $\bar{x} = 1.8$)

Equation $\bar{x} = r\,\frac{\sigma_x}{\sigma_y} \times y$ gives the relationship between x and y in *deviation form*. That is to say, it gives the most probable

deviation of an X-measure from M_X corresponding to a known *deviation* in the Y-measure from M_Y.

Although both of the regression equations given above involve x and y, the two equations cannot be used interchangeably — neither can be used to predict *both* x and y. This is an important fact which the reader must understand clearly and constantly bear in mind. The first regression equation $\bar{y} = r\frac{\sigma_y}{\sigma_x} \times x$ can be used *only* when y is to be predicted from a given x (when y is the "dependent" variable) *; while the second equation $\bar{x} = r\frac{\sigma_x}{\sigma_y} \times y$ can be used *only* when x is to be predicted from a known y (when x is the "dependent" variable). There are always *two* regression equations in a correlation table, the one through the means of the columns and the other through the means of the rows, unless the correlation is 1.00 or − 1.00. When $r = 1.00$, $\bar{y} = r\frac{\sigma_y}{\sigma_x} \times x$ becomes $\bar{y} = \frac{\sigma_y}{\sigma_x} \times x$ or $\bar{y}\sigma_x = x\sigma_y$. Also, when $r = 1.00$, $\bar{x} = r\frac{\sigma_x}{\sigma_y} \times y$ becomes $\bar{x} = \frac{\sigma_x}{\sigma_y} \times y$ or $\bar{x}\sigma_y = y\sigma_x$. In short, when the correlation is perfect (± 1.00), the two equations are identical and the two regression lines coincide. To illustrate this situation, suppose that the correlation between height and weight in Figure 52 were perfect. The first regression equation would then be $\bar{y} = 1.00 \times \frac{2.62}{15.54}x$ or $\bar{y} = .17x$, and the second, $\bar{x} = 1.00 \times \frac{15.54}{2.62}y$, or $\bar{x} = 5.93y$. Algebraically, the equation $x = 5.93y$ is equal to $y = .17x$; for if we put $x = \frac{y}{.17}$, $x = 5.93y$. When $r = \pm 1.00$ there is only *one* equation and a *single* regression line. Moreover, if $r = \pm 1.00$, and in addition $\sigma_x = \sigma_y$, the single regression line makes an angle of 45° or 135° with the horizontal axis, since $y = \pm x$.

* The dependent variable takes its value from the other (independent) variable in the equation. For example, in the equation $y = 3x^2 + 5x - 10$, y "depends" for its value upon x; hence y is the dependent variable.

3. Plotting the Regression Lines in a Correlation Table *

In Figure 52, the coördinate axes have been drawn in on the correlation table through the means of the X- and Y-distribu-

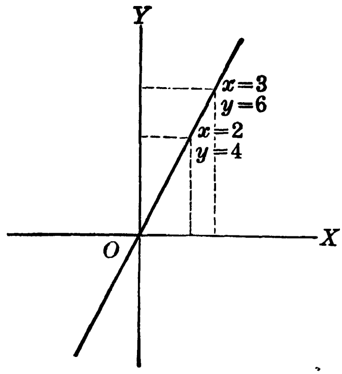


FIG. 53. Plot of the Straight Line, $y = 2x$.

* A brief review of the equation of a straight line, and of the method of plotting a simple linear equation is given here in order to simplify the plotting of the regression equations.

In Figure 53, let X and Y be coördinate axes, or axes of reference. Now suppose that we are given the equation $y = 2x$ and are required to represent the relation between x and y graphically. To do this we assign values to x in the equation and compute the corresponding values of y. When $x = 2$, for example, $y = 2 \times 2$ or 4; when $x = 3$, $y = 2 \times 3$ or 6. In the same way, given any x-value we can compute the value of y which will "satisfy" the equation, that is, make the left side equal to the right. If the series of x and y values found from the equation are plotted on the diagram with respect to the X- and Y-coördinates (as in Fig. 53) they will be found to fall along a straight line. This straight line pictures the relation $y = 2x$. It goes through the origin, since when $x = 0$, $y = 0$. The equation $y = 2x$ represents, then, a straight line which passes through the origin; and the relation of its coördinates (points lying along the line) is such that $\frac{y}{x}$, called the *slope* of the line, is always equal to 2.

The general equation of any straight line which passes through the origin may be written $y = mx$, where m is the slope of the line. If we replace m in the general formula by $r\frac{\sigma_y}{\sigma_x}$ it is clear that the regression line in *deviation form*, namely, $y = r\frac{\sigma_y}{\sigma_x}x$, is simply the equation of a straight line which goes through the origin. For the same reason, when the general equation of a straight line through the origin is written $x = my$, $x = r\frac{\sigma_x}{\sigma_y}y$ is also seen to be a straight line through the origin, its slope being $r\frac{\sigma_x}{\sigma_y}$.

tions. The vertical axis is drawn through 136.3 pounds (M_{wt}), and the horizontal axis through 66.5 inches (M_{ht}). These axes intersect close to the center of the chart. Equations (54) and (55) define straight lines which pass through the origin or point of intersection of these coördinate axes. For this reason, it is a comparatively simple task to plot in our regression lines on the correlation chart with reference to the given coördinate axes

Correlation charts are usually laid out with equal distances representing the X and Y class-intervals (the printed correlation charts are always so constructed) although the intervals expressed in terms of the variables themselves may be, and often are, unequal and incommensurable. This is true in Figure 52. In this diagram, the intervals in X and Y appear to be equal, although the actual interval for height is 2 inches, and the actual interval for weight is 10 pounds. Because of this difference in interval-length in the two variables it is very important that we express σ_x and σ_y in our regression equations in *class-interval units* before plotting the regression lines on the chart. Otherwise we must equate our X and Y intervals by laying out our diagram in such a way as to make the X-interval five times the Y-interval. This latter method of equating intervals is impractical, and is rarely used, since all we need do in order to use correlation charts drawn up with equal intervals is to express σ_x and σ_y in formulas (54) and (55) in units of interval. When this is done, and the interval, *not* the score, is the unit, the first regression equation becomes

$$y = .60\,\frac{1.31}{1.55}\,x \quad \text{or} \quad y = .51x$$

and the second

$$x = .60\,\frac{1.55}{1.31}\,y \quad \text{or} \quad x = .71\,y$$

Since each regression line goes through the origin, only one other point (besides the origin) is needed in order to determine its course. In the first regression equation, if $x = 10$, $y = 5.1$; and the two points (0, 0) and (10, 5.1) locate the line. In the second regression equation, if $y = 10$, $x = 7.1$; and the two

points (0, 0) and (7.1, 10) determine the second line. In plotting points on a diagram any convenient scale may be employed. A millimeter rule is useful.

It is important for the reader to remember that when the two σ's are expressed in interval units, regression equations do *not* give the relationship between the X and Y score deviations. These special forms of the regression equations should not be used except when plotting the equations on a correlation chart. Whenever the most probable deviation in the one variable corresponding to a known deviation in the other is wanted, formulas (54) and (55), in which the σ's are expressed in *score units*, must be employed.

4. The Regression Equations in Score Form

In the last sections it was pointed out that formulas (54) and (55) give the equations of the regression lines in deviation form — that values of x and y substituted in these equations are deviations from the means of the X and Y distributions, and are not scores. While the equations in *deviation form* are actually all that one needs in order to pass from one variable to another, it is decidedly convenient to be able to estimate an individual's *actual score* in Y, say, directly from the score in X without first converting the X-score into a deviation from M_X. This can be done by using the *score form* of the regression equations. The conversion of deviation form to score form is made as follows: Denoting the mean of the Y's by M_Y and any Y-score simply by Y, we may write the deviation of any individual from the mean as $Y - M_Y$ or, in general, $y = Y - M_Y$. In the same way, $x = X - M_X$ when x is the deviation of any X-score from the mean X. If we substitute $Y - M_Y$ for y, and $X - M_X$ for x, in formulas (54) and (55), the two regression equations become

$$Y - M_Y = r\frac{\sigma_y}{\sigma_x}(X - M_X) \text{ or } \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - M_X) + M_Y \quad (56)$$

and

$$X - M_X = r\frac{\sigma_x}{\sigma_y}(Y - M_Y) \text{ or } \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - M_Y) + M_X \quad (57)$$

(*regression equations of* Y *on* X *and* X *on* Y *in score form*)

These two equations are now said to be in *score form,* since the X and Y in both equations represent *actual scores, and not deviations from the means* of the two distributions.

If we substitute in (56) the values of M_Y, r, σ_y, σ_x and M_X obtained from Figure 52, the regression of height on weight in score form becomes

$$\bar{Y} = .60 \times \frac{2.62}{15.54}(X - 136.3) + 66.5$$

or upon reduction

$$\bar{Y} = .10X + 52.9$$

To illustrate the use of this equation, suppose that a man in our group weighs 160 pounds and we wish to estimate his most probable height. Substituting 160 for X in the equation, $\bar{Y} = 69$ inches; and accordingly, the most probable height of a man who weighs 160 pounds is 69 inches.

If the problem is to predict weight instead of height, we must use the second regression equation, formula (57). Substituting for M_X, r, σ_x, σ_y, and M_Y in (57) we have

$$\bar{X} = 60 \times \frac{15.54}{2.62}(Y - 66.5) + 136.3$$

or

$$\bar{X} = 3.56Y - 100.4$$

Now if a man is 71 inches tall, we find, on replacing Y by 71 in the equation, that $\bar{X} = 152.4$. Hence the most probable weight of a man who is 71 inches tall is about $152\frac{1}{2}$ pounds.

5. The Meaning of a "Prediction" from the Regression Equation

It may seem strange, perhaps, to talk of "predicting" a man's height from his weight, when the heights and weights of the 120 men in our group are already known. When we have measures of both height and weight it is unnecessary to estimate one from the other. But suppose that all we know about a given individual is his weight and the fact that he falls within the age-range of our group of 120 men. Since we know the

correlation between height and weight to be .60, it is possible from the regression equation to predict the most probable height of our subject in lieu of actually measuring him. Furthermore, the regression equation may be employed to estimate the height of *any* man in the population from which our group is chosen, provided our sample is an unbiased selection from the larger group. A regression equation holds, of course, only for the population from which the sample group was drawn We cannot estimate the heights of children or of women from a regression equation which describes the relationship between height and weight in men between the ages of eighteen and twenty-five years (the age-range of the students in our group). Conversely, we cannot expect a regression equation established for elementary school children to hold for older groups.

Height and weight, since they are both easily measured, perhaps do not demonstrate the value of the regression equation so clearly as do other and more complex traits. These variables were chosen for our "model" problem because they are objective and observable and their meaning is definite. Let us now consider a problem of more direct psychological interest. Suppose that in a group of 300 high-school children of nearly the same age, the correlation between group test scores obtained at the beginning of the school year and average grades made in the first year of high school is .60. Now if we administer the group test to a child who enters school the next year, it is possible from his score to estimate his probable scholastic performance by means of the regression equation between test score and grades obtained from the previous year's class. Forecasts of this sort are useful in educational prognosis and guidance.* The same is true of vocational guidance; we are often able to predict from a test battery the probable success of an individual who contemplates entering a certain trade or profession.† Advice on such a basis is measurably better than subjective judgment.

* Edgerton, H. A., *Academic Prognosis in the University*, Educational Psychology Monographs, 27 (1930).

† Stead, W. H., and Shartle, C. L., *Occupational Counseling Techniques* (1940).

II. The Reliability of Predictions

1. The Standard Error of Estimate

The values of X and Y "predicted" from regression equations have been constantly referred to as being the "most probable" values of the one variable accompanying the given value of the other. In order to show just how probable such estimates are it is necessary that we calculate their standard errors of estimate. The accuracy with which we are able to predict Y-scores from equation (56) is given by the formula

$$\sigma_{(\text{est. } Y)} = \sigma_y\sqrt{1 - r^2} \tag{58}$$

[*standard error of a* Y-*score predicted from equation* (56)] *

in which σ_y is the σ of the Y distribution, and r is the coefficient of correlation. The subscript "est." is used to distinguish this standard error from the σ of the distribution, the σ_M, etc.

From formula (56) we have calculated the most probable height of a man weighing 160 pounds to be 69 inches. The reliability of this prediction is obtained by substituting $\sigma_{(ht)}$ and r in formula (58) to find

$$\sigma_{(\text{est. } Y)} = 2.62\sqrt{1 - .60^2} = 2.1 \text{ inches}$$

We now say that the most probable height of a man weighing 160 pounds is 69 inches with a $\sigma_{(\text{est.})}$ of 2.1 inches; and that the chances are about two in three that our prediction does not miss the man's actual height by more than $\pm$ 2.1 inches. We may feel quite certain that the estimated height of this man does not miss his true height by more than $\pm 3\sigma_{(\text{est.})}$ or by more than $\pm$ 6.3 inches.

The degree of accuracy with which X-scores can be predicted from (57) is given by the formula

$$\sigma_{(\text{est. } X)} = \sigma_x\sqrt{1 - r^2} \tag{59}$$

[*standard error of an* X-*score predicted from equation* (57)]

* The probable error of estimate is $PE_{(\text{est. } Y)} = .6745\ \sigma_y\sqrt{1 - r^2}$

in which σ_x is the σ of the X distribution, and r is the coefficient of correlation.

We found on page 318 that the most probable weight of a man in our group who is 71 inches tall is 152.4 pounds. The $\sigma_{(est.)}$ of this prediction from (59) is

$$\sigma_{(est.\ X)} = 15.54\sqrt{1 - .60^2} = 12.4 \text{ pounds}$$

and the most probable weight of *any* man 71 inches tall, in our group or in the population from which it is drawn, is 152.4 pounds with a $\sigma_{(est.)}$ of 12.4 pounds. The chances, therefore, are about two in three that our prediction does not miss our man's true weight by more than $\pm$ 12.4 pounds.

2. The Accuracy of Individual Predictions from Regression Equations

The formulas for $\sigma_{(est.)}$ measure the error made in taking predicted, instead of actual, X and Y measures. If $r = 1.00$, $\sqrt{1 - r^2}$ is 0, and $\sigma_{(est.)}$ is zero — there is *no* error of estimate and each person's measurement is predicted exactly. On the other hand, when $r = .00$, $\sqrt{1 - r^2} = 1.00$, and the error of estimate is equal to the σ of the distribution into which prediction is made. When this last situation occurs, the regression equation is of no value in enabling us the better to predict scores, as each person's most probable score (e.g., X) is simply the mean (i.e., M_X). When $r = .00$ all that we can say definitely is that a subject's score lies *somewhere* in the distribution of Y's or X's. But just where we cannot tell, since our *SE* of estimate equals the *SD* of the test.

It is clear from formulas (58) and (59) that the accuracy of prediction from a regression equation depends directly upon the σ's of the two distributions (σ_y or σ_x) and upon the degree of correlation between the two sets of measures. If the variability (σ_y) of Y is small, and the correlation between Y and X high (e.g., .90), values of Y can be predicted from known values of X with a comparatively high degree of accuracy. However, when the variability of a test is large, or the correlation low (or when

both conditions obtain), prediction from regression equations becomes so unreliable as to be almost valueless. Even when the correlation is fairly high, forecasts will often have an uncomfortably large error of estimate. Thus we have seen that in spite of the $r = .60$ between height and weight (Fig. 52), our forecast of a man's weight, knowing his height, has a $\sigma_{(\text{est. } X)}$ of about 12 pounds (p. 321). Prediction of height from weight is somewhat better than prediction of weight from height. Predicted heights will, in two-thirds of the cases, be in error by not more than 2 inches. An example in which high correlation offsets fairly large variability, permitting reasonably accurate forecasts, is given later in Figure 54.

When an investigator uses the regression equations for purposes of prediction, he should always give the $\sigma_{(\text{est.})}$ of his estimated scores. The value of a forecast depends, first of all, upon the size of the error of estimate; but it also depends upon the units of measurement, and upon the purposes for which the prediction is made (p. 333).

3. The Accuracy of Group Predictions

We have seen in (2) above that the standard error of a predicted score ($\sigma_{(\text{est.})}$) may often be uncomfortably large. Only when $r = 1.00$ is $\sqrt{1 - r^2} = .00$, and only then can an estimate be made without error. The correlation coefficient must be .87 before $\sqrt{1 - r^2}$ is .50, i.e., before the standard error of estimate is reduced 50% below the σ of the test. Obviously, unless r is quite large (larger than we usually get in practice) the regression equation is of little aid in forecasting with reasonable accuracy what a given individual may be expected to do (p. 334). This has led many to discount unwisely the value of correlation in prediction and to conclude that the calculation of r is not worth the trouble.

Fortunately correlation makes out better in forecasting the performance of *groups* than in predicting the most likely achievement of a given *individual.* In forecasting achievement the psychologist is in much the same position as the insurance stat-

istician or actuary. The actuary cannot tell how long John Smith, aged twenty, will live. But from his tables, he can tell quite accurately how many of 10,000 men now aged twenty will live to be thirty, forty, or fifty years old. In the same way, the psychologist may be quite uncertain concerning the performance of a given individual. But knowing the correlation between a test (or test battery) and some criterion of performance, he can forecast often with considerable accuracy the probable performance of various groups chosen from his distribution of test scores. The degree of accuracy in such predictions depends upon the size of the correlation coefficient.

To illustrate "actuarial" prediction in psychology, suppose that 70% of a freshman class of 400 men achieve grades in their college work above the minimum passing mark and hence are regarded as "satisfactory" students. Suppose, further, that the correlation between a standard intelligence test and freshman performance is .50. Now if we had selected the *upper half* of our group (i.e., the 200 students who performed *best* on the intelligence test) at the beginning of the term, how many of these 200 would have been "satisfactory," i.e., in the upper 70% of the "grades" distribution? From Table 50 it can easily be read that 84% of our 200 selected freshmen (i.e., 168) should be found in the satisfactory group with respect to grades. The entry .84 is found in column .50 (percentage of test distribution chosen) opposite the correlation of .50. This result should be compared with the 70% (i.e., 140) who might be expected to fall in the satisfactory group when selection is by "guess," without knowledge of the correlation. This entry is in column .50 opposite the r of .00.

The probable performance of other and smaller groups chosen from our test distribution can be estimated with much greater accuracy from Table 50. We know, for example, that 91% of the best 20% of our students (roughly, seventy-three in the first eighty) can be expected to prove satisfactory in terms of our criterion (i.e., being located in the upper 70% of the grade distribution). Read the entry .91 in column 20 opposite $r = .50$.

TABLE 50 *

PROPORTION OF STUDENTS CONSIDERED SATISFACTORY IN TERMS OF GRADES = .70

Selection Ratio: Proportion Selected on Basis of Tests

r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.70	.70	.70	.70	.70	.70	.70	.70	.70	.70	.70
.05	.73	.73	.72	.72	.72	.71	.71	.71	.71	.70	.70
.10	.77	.76	.75	.74	.73	.73	.72	.72	.71	.71	.70
.15	.80	.79	.77	.76	.75	.74	.73	.73	.72	.71	.71
.20	.83	.81	.79	.78	.77	.76	.75	.74	.73	.71	.71
.25	.86	.84	.81	.80	.78	.77	.76	.75	.73	.72	.71
.30	.88	.86	.84	.82	.80	.78	.77	.75	.74	.72	.71
.35	.91	.89	.86	.83	.82	.80	.78	.76	.75	.73	.71
.40	.93	.91	.88	.85	.83	.81	.79	.77	.75	.73	.72
.45	.94	.93	.90	.87	.85	.83	.81	.78	.76	.73	.72
.50	.96	.94	.91	.89	.87	.84	.82	.80	.77	.74	.72
.55	.97	.96	.93	.91	.88	.86	.83	.81	.78	.74	.72
.60	.98	.97	.95	.92	.90	.87	.85	.82	.79	.75	.73
.65	.99	.98	.96	.94	.92	.89	.86	.83	.80	.75	.73
.70	1.00	.99	.97	.96	.93	.91	.88	.84	.80	.76	.73
.75	1.00	1.00	.98	.97	.95	.92	.89	.86	.81	.76	.73
.80	1.00	1.00	.99	.98	.97	.94	.91	.87	.82	.77	.73
.85	1.00	1.00	1.00	.99	.98	.96	.93	.89	.84	.77	.74
.90	1.00	1.00	1.00	1.00	.99	.98	.95	.91	.85	.78	.74
.95	1.00	1.00	1.00	1.00	1.00	.99	.98	.94	.86	.78	.74
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.88	.78	.74

If the correlation of the intelligence test and school grades had been .60 instead of .50, 87% (or 174 in 200) of the "best half" according to the test would have been satisfactory students; and 95% of the "best" 20% on the test should be satisfactory students. These forecasts are to be compared with 70%, the estimate when $r = .00$. It is clear that a knowledge of the correlation greatly improves the estimate, and the larger the r the better the forecast.

Table 50 is a small part of a larger table in which "proportions considered satisfactory in achievement" range from .05 to .95.†

* Taylor, H. C., and Russell, J. T., "The Relationships of Validity Coefficients to the Practical Effectiveness of Tests in Selection: Discussion and Tables," *Journal of Applied Psychology*, 23 (1939), 565–578.

† Taylor, H. C., and Russell, J. T., *op. cit.*

The correlation between test score and performance ranges from .00 to 1.00. These tables are strictly accurate only when the distributions are normal both in the test and in the criteria of performance. They may be used with considerable confidence, however, when the distributions are approximately normal, especially when the N's are large; and in any case they furnish useful approximations.

Forecasting tables have considerable value in selecting personnel for business or other vocations. First, we must determine what proportion of a given group of workers is to be considered "successful." With this information in hand and knowing the correlation between our test battery and performance in the given activity, we may forecast the probable success of groups of new applicants from their test scores. Assume, for example, that 70% of a group of factory workers are regarded as "acceptable workers," acceptability having been determined from ratings by foremen, number of pieces done in a given time, or time taken to complete certain standard jobs. Assume, further, that a test battery has a correlation of .45 with worker-performance. Then if we select the best twenty out of 100 applicants ("best" according to our tests), we find from Table 50 that 90% of this number or eighteen should be acceptable workers. If we had had no test and had simply selected the *first* twenty applicants to appear — or *any* twenty — 70% or fourteen should be acceptable Use of the tests improves our forecast 30%; and the more stringent the criterion of acceptability the greater the improvement in forecast made by the tests.

III. The Effect of Variability of Measures upon the Size of r

Suppose that the correlation between two tests in a small group of fifty sixth-grade children has been found to be .50. How will this correlation compare with that between the same tests in a larger group of greater range, e.g., a group of 200 children in the sixth grade or 200 children spread over grades

six, seven, and eight? More generally, knowing the correlation between two tests in a group of narrow range, can we predict the probable correlation in a group of wider range?

The problem of the effect upon r of the "range of talent" (size of σ_x and σ_y) within the group being studied often arises in correlational work. It becomes important, for example, when one wishes to go beyond the correlation obtained in the sample with which he is working and generalize (estimate the r) for a group of wider range; or when r's between the same tests obtained in different ranges are to be compared. A formula for estimating the correlation between two tests in a heterogeneous group when we know the correlation between the tests in a homogeneous group may be developed in the following way: Let $\sigma_{(\text{est. } Y_s)}$ be the standard error of estimate in a group somewhat curtailed in variability or in range of talent; and $\sigma_{(\text{est. } Y_l)}$ be the standard error of estimate in a larger group less restricted in variability. (Y is the *dependent* variable, p. 314.) Then, on the assumption that our tests are as effective in the *wide* as in the *narrow* range, $\sigma_{(\text{est. } Y_l)} = \sigma_{(\text{est. } Y_s)}$, or, by formula (58), page 320,

$$\sigma_{y_s}\sqrt{1 - r^2_{x_s y_s}} = \sigma_{y_l}\sqrt{1 - r^2_{x_l y_l}}$$

and

$$\frac{\sigma_{y_s}}{\sigma_{y_l}} = \frac{\sqrt{1 - r^2_{x_l y_l}}}{\sqrt{1 - r^2_{x_s y_s}}} \tag{60}$$

(formula for estimating correlation in a wide *range from a knowledge of the correlation in a* narrow *range)*

in which σ_{y_s} is the standard deviation of Y in the small group, or in the curtailed range; σ_{y_l} is the standard deviation of Y in the large group, or in the uncurtailed range; $r_{x_s y_s}$ = the correlation in the small group, and $r_{x_l y_l}$ = the correlation in the large group.

To illustrate formula (60), suppose that in a small group $\sigma_{y_s} = 10$ and $r_{x_s y_s}$ is .50. What would the r between the same two tests probably be in a group in which $\sigma_{y_l} = 15$: in which

σ_{y_l} is 50% larger than σ_{y_s}? Substituting $\sigma_{y_s} = 10$, $\sigma_{y_l} = 15$, and $r_{x_s y_s} = .50$ in (60), we have

$$\frac{10}{15} = \frac{\sqrt{1 - r^2_{x_l y_l}}}{\sqrt{1 - .25}}\,.$$

Squaring both sides of this equation, and solving, we have $r_{x_l y_l} = .82$. The r of .50 in the narrow range becomes an r of .82 in the wider range. It is clear from this example that direct comparison of r's is not valid when the variabilities (σ's) within the groups from which the r's were computed are quite different.

If X and not Y is the dependent variable, formula (60) becomes

$$\frac{\sigma_{x_s}}{\sigma_{x_l}} = \frac{\sqrt{1 - r^2_{x_l y_l}}}{\sqrt{1 - r^2_{x_s y_s}}} \tag{61}$$

(formula for estimating correlation in a wide range from a knowledge of the correlation in a narrow range)

Formulas (60) and (61) are open to the objection that each takes account of only *one* distribution in estimating the probable increase in r with increase in range of talent. If, however, the increase in σ_y as the group becomes more heterogeneous is accompanied by a proportional increase in σ_x (or vice versa), formulas (60) and (61) will give accurate estimates. Experimental trial of these formulas has yielded results closely in accord with theoretical expectation.*

IV. The Solution of a Second Correlation Problem

The solution of a second correlation problem will be found in Figure 54. The purpose of another "model" is to strengthen the reader's grasp of correlational techniques by having him work straight through the process of calculating r and the regression equations upon a new set of data. A student often fails to relate the various aspects of a correlational problem when these are presented in piecemeal fashion.

* Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases* (1940), pp. 208–212.

1. Calculation of r

Our first problem in Figure 54 is to find the correlation between the I.Q.'s achieved by 190 children of the same — or approximately the same — chronological age who have taken an intelligence examination upon two occasions separated by a six months' interval. The correlation table has been constructed from a scattergram, as described on page 275. The test given first is the X-variable, and the test given second is the Y-variable. The calculation of the two means, and of c_x, c_y, σ_x, and σ_y covers familiar ground, is given in detail on the chart, and need not be repeated here.

The product-deviations in the $\Sigma x'y'$ column have been taken from column 100–104 (column containing the AM_X) and from row 105–109 (row containing the AM_Y). The entries in the $\Sigma x'y'$ column have been calculated by the shorter method described on page 286; that is, each cell entry in a given row has been multiplied *first* by its x-deviation (x') and the sum of these deviations entered in the column $\Sigma x'$. The $\Sigma x'$ entries were then "weighted" once for all by the y' of the whole row. To illustrate, in the first row reading from left to right $(1 \times 5) + (1 \times 6)$ or 11 is the $\Sigma x'$ entry. The x''s are 5 and 6, respectively, and may be read from the x' row at the bottom of the correlation table. Since the common y' is 5, the final $\Sigma x'y'$ entry is 55. Again in the seventh row reading down from the top of the diagram $(5 \times -3) + (3 \times -2) + (7 \times -1) + (16 \times 0) + (2 \times 1) + (4 \times 2)$ or -18 makes up the $\Sigma x'$ entry. The y' of this row is -1, and the final $\Sigma x'y'$ entry is 18. To take still a third example, in the eleventh row from the top of the diagram, $(1 \times -5) + (3 \times -4) + (1 \times -3) + (2 \times -2)$ or -24 is the $\Sigma x'$ entry. The common y' is -5 and the $\Sigma x'y'$ entry is 120.

Three checks of the calculations (see p. 283), upon which r, σ_x and σ_y are based, are given in Figure 54. Note that $\Sigma fx' = \Sigma x'$; and that, when the $\Sigma x'y'$'s are recalculated, at the bottom of the chart, $\Sigma fy' = \Sigma y'$, and the two determinations of $\Sigma x'y'$ are equal. When the $\Sigma x'y'$'s have been checked, the cal-

Second Test (Y) \ First Test (X)	75-79	80-84	85-89	90-94	95-99	100-104	105-109	110-114	115-119	120-124	125-129	130-134	f_y	y'	fy'	fy'^2	$\Sigma x'$	$\Sigma x'y'$
130-134											1	1	2	5	10	50	11	55
125-129								1	2	2	2	1	8	4	32	128	32	128
120-124							3	2	3	2	1	1	12	3	36	108	35	105
115-119							1	4	4	2	2		13	2	26	52	39	78
110-114				1	1	2	3	2	1	2	1		13	1	13 (117)	13	20	20
105-109			2	6	3	5	3	3	4	2			28	0		0	8	0
100-104			5	3	7	16	2	4					37	−1	−37	37	−18	18
95-99		3	6	4	5	3	3	4					28	−2	−56	112	−32	64
90-94		4	2	5	2	4	1	2					20	−3	−60	180	−29	87
85-89	1	8	1	3	5	2							20	−4	−80	320	−51	204
80-84	1	3	1	2									7	−5	−35	175	−24	120
75-79	1		1										2	−6	−12 (−280)	72	−8	48
f_x	3	18	18	24	23	32	16	22	14	10	7	3	190		−163	1247	−17	927
x'	−5	−4	−3	−2	−1	0	1	2	3	4	5	6						
fx'	−15	−72	−54	−48	−23 (−212)		16	44	42	40	35	18 (195)	= −17					
fx'^2	75	288	162	96	23		16	88	126	160	175	108	= 1317					
$\Sigma y'$	−15	−65	−38	−47	−42	−40	3	2	26	20	21	12	= −163					
$\Sigma x'y'$	75	260	114	94	42	0	3	4	78	80	105	72	= 927					

check

check

check

$c_y = -\frac{163}{190} = -.86$

$c^2_y = .74$

$c_y i = -.86 \times 5 = -4.30$

$M_Y = 107 - 4.30 = 102.7$

$c_x = -\frac{17}{190} = -.09$

$c^2_x = .008$

$c_x i = -.09 \times 5 = -.45$

$M_X = 102 - .45 = 101.55$

SE's of Estimate

$\sigma_{(est.Y)} = 12.05\sqrt{1 - .76^2}$
$= 7.83$

$\sigma_{(est.X)} = 13.15\sqrt{1 - .76^2}$
$= 8.55$

$\sigma_y = \sqrt{\frac{1247}{190} - .74} \times 5$
$= 2.41 \times 5$
$= 12.05$

$\sigma_x = \sqrt{\frac{1317}{190} - .008} \times 5$
$= 2.63 \times 5$
$= 13.15$

$r = \frac{\frac{927}{190} - (-.86 \times -.09)}{2.41 \times 2.63}$
$= .76$

Regression Equations

$\bar{y} = .76 \times \frac{12.05}{13.15} x = .69x$

$\bar{x} = .76 \times \frac{13.15}{12.05} y = .83y$

$Y - 102.7 = .69(X - 101.6)$
$\overline{Y} = .69X + 32.6$

$X - 101.6 = .83(Y - 102.7)$
$\overline{X} = .83Y + 16.4$

FIG. 54. Calculation of the Correlation between the I.Q.'s Achieved by 190 Children of the Same C.A. upon Two Forms of an Individual Intelligence Examination.

culation of r by formula (44) is a matter of substitution. Note carefully that c_x, c_y, σ_x, σ_y are all left in *units of class-interval* in the formula for r (p. 288).

2. Calculation of the Regression Equations and the *SE*'s of Estimate

The regression equations in deviation form are given on the chart and the two lines which these equations represent have been plotted on the diagram. Note that these equations may be plotted as they stand, since the class-interval is the same for X and for Y (p. 316). In the routine solution of a correlational problem it is not strictly necessary to plot the regression lines on the chart. These lines are often of value, however, in indicating whether the means of the X- and Y-arrays can be represented by straight lines, that is, whether regression is "linear." If the relationship between X and Y is not linear, other methods of calculating the correlation must be employed (p. 365).

The standard errors of estimate, shown in Figure 54, are 7.83 and 8.55 depending upon whether the prediction is of Y from X or X from Y. All I.Q.'s predicted on the Y-test from X may be considered to have the same error of estimate,* and similarly for all predictions of X from Y.

Errors of estimate are most often used to give the reliability of specific predicted measures. But they also have a more general interpretation. Thus a $\sigma_{(\text{est. } Y)}$ of 7.83 points means that two-thirds of the I.Q.'s in test Y missed perfect correspondence with the I.Q.'s in test X by $\pm$ 7.83 points or *less*, while the other one-third missed complete agreement by *more* than $\pm$ 7.83 points. Stated differently, we may say that 68% of the I.Q.'s predicted on test Y from test X may be expected to differ from their actual values by not more than $\pm$ 7.83 points, while the remaining 32% may be expected to differ from their actual values by more than $\pm$ 7.83 points.

* See, however, Terman, L. M., and Merrill, M. A., *Measuring Intelligence* (1937), pp. 44–47, where the *SE*'s of estimate have been computed for various IQ levels.

3. The "Regression Effect" in Prediction

Predicted scores tend to "move in" toward the mean of the distribution into which prediction is made (p. 311). This so-called regression effect has often been noted by investigators and is always present when correlation is less than $\pm$ 1.00.* The regression phenomenon can be clearly seen in the following illustrations: From the regression equation $\bar{Y} = .69X + 32.6$ (Fig. 54) it is clear that a child who earns an I.Q. of 130 on the first test (X) will most probably earn an I.Q. of 122 on the second test (Y); while a child who earns an I.Q. of 120 in X will most probably score 115 in Y. In both of these illustrations the predicted Y-test I.Q. is smaller than the first or X-test I.Q. Put differently, the second I.Q. has *regressed* or moved down toward the mean of test Y, i.e., toward 102.7. The same effect occurs when the I.Q. on the X-test is *below* its mean; the tendency now is for the predicted score in Y to move *up* toward its mean. Again, from the equation $\bar{Y} = .69X + 32.6$, we find that if a child earns an I.Q. of 70 on the X-test his most likely score on the second test (Y) is 81; while an I.Q. of 80 on the first test forecasts an I.Q. of 88 on the second. Both of these predicted I.Q.'s have moved up nearer to the mean 102.7 (M_y).

The tendency for all scores predicted from a regression equation to pull in — down or up — toward the mean, can best be seen as a general phenomenon if the regression equation is written in standard-score form. Given

$$\bar{y} = r\frac{\sigma_y}{\sigma_x} \times x \qquad \text{(54) p. 312}$$

if we divide both sides of this equation by σ_y and write σ_x under x, we have

$$\frac{\bar{y}}{\sigma_y} = r\frac{x}{\sigma_x} \text{ or } \bar{z}_y = rz_x \qquad (62)$$

(*regression equation when scores in* X *and* Y *are expressed as* z *or standard-scores*)

* Thorndike, R. L., "Regression Fallacies in the Matched Groups Experiment," *Psychometrika*, 7 (1942), 85–102.

In the problem in Figure 54, $\bar{z}_y = .76z_x$. If z_x is $\pm$ 1.00σ, or $\pm$ 2.00σ, or $\pm$ 3.00σ from M_x, $\bar{z}_y$ will be $\pm$ .76σ, $\pm$ 1.52σ, or $\pm$ 2.28σ from M_y. That is to say, *any* score above or below the mean of X forecasts a Y-score somewhat *closer* to the mean of Y.

In studying the relation of height in parent and offspring, Galton (p. 311) interpreted the phenomenon of regression to the mean to be a provision of nature designed to protect the race from extremes This same effect occurs, however, in any correlation table in which r is less than $\pm$ 1.00, and need not be explained in biological terms. The I.Q.'s of a group of very bright children, for instance, will tend upon retest to move *downward* toward 100, the mean of the group; while the I.Q.'s of a group of dull children will tend upon retest to move *upward* toward 100.

V. The Interpretation of the Coefficient of Correlation

When should a coefficient of correlation be called "high," when "medium," and when "low"? Does an r of .40 between two tests indicate "marked" or "low" relationship? How high should an r be in order to permit accurate prediction from one variable to another? Can an r of .50, say, be interpreted with respect to "overlap" of determining factors in the two variables correlated? Questions like these, all of which are concerned with the *significance* or *meaning* of the relationship expressed by a correlation coefficient constantly arise in problems involving mental measurement, and their implications must be understood before we can effectively employ the correlational method.

The value of r as a measure of correspondence may be profitably considered from two points of view.* In the first place, r's are computed in order to determine whether there is *any* correlation (over and above chance) between two variables; and in the second place, r's are computed in order to determine

* Barr, A. S., "The Coefficient of Correlation," *Journal of Educational Research*, 23 (1931), 55–60.

the *degree* or closeness of relationship when some association is known, or is assumed, to exist. The question, "Is there *any* correlation between brain weight and intelligence?", voices the first objective. And the question, "*How significant* is the correlation between high-school grades and first-year performance in college?", expresses the second. The problem of when an obtained r denotes significant relationship has already been considered on page 297. This section is concerned mainly with the second problem, namely, the evaluation — with respect to *degree* of relationship — of an obtained coefficient. The questions at the beginning of the paragraph above all bear upon this topic.

1. The Interpretation of *r* in Terms of Verbal Description

It is customary in mental measurement to describe the correlation between two tests in a general way as high, marked or substantial, low or negligible. While the descriptive label applied will vary somewhat in meaning with the author using it, there is fairly good agreement among workers with psychological and educational tests that an

r from .00 to $\pm$ .20 denotes indifferent or negligible relationship;
r from $\pm$ .20 to $\pm$ .40 denotes low correlation; present but slight;
r from $\pm$ .40 to $\pm$ .70 denotes substantial or marked relationship;
r from $\pm$ .70 to $\pm$ 1.00 denotes high to very high relationship.

This classification is broad and somewhat tentative, and can only be accepted as a general guide with certain reservations. Thus a coefficient of correlation must always be judged with regard to

(1) the nature of variables with which we are dealing;
(2) the significance of the cœfficient;
(3) the size and variability of the group (p. 325);
(4) the reliability cœfficients of the tests used (p. 380);
(5) the purpose for which the r was computed.

To consider, first, the matter of the variables being correlated, an r of .30 between height and intelligence, or between head measurements and mechanical ability would be regarded as

important although it is rather low, since correlations between physical and mental functions are usually much lower — often zero. On the other hand, the correlation must be .70 or more between measures of general intelligence and school grades or between achievement in English and in history to be considered high, since *r*'s in this field usually run from .40 to .60. Resemblances of parents and offspring with respect to physical and mental traits are expressed by *r*'s of .35 to .55; and, accordingly, an *r* of .60 would be high.* By contrast, the reliability of a standard intelligence test is ordinarily much higher than .60, and the self-correlation of such a test must be .85 to .95 to be regarded as high. In the field of vocational testing, the *r*'s between test batteries and measures of aptitude represented by various criteria rarely rise above .50 †; and *r*'s above this figure would be considered exceptionally promising.

Correlation coefficients must be evaluated also with due regard to the reliabilities (p. 380) of the two tests concerned. Because of chance errors, an obtained *r* is always less than its "corrected" value (p. 396) and hence, in a sense, is a minimum measure of the relationship present. The effect upon an *r* of the size and variability of the group is discussed elsewhere (p. 325), and a formula for estimating such effect provided. The *purpose* for which the correlation has been computed is important. The *r* which is to be employed in predicting the scores of *individuals* from one test to another, for instance, should be much higher than the *r*, the purpose of which is to provide forecasts of the achievement of selected groups (p. 322).

In summary, a correlation coefficient is always to be judged with reference to the circumstances under which it was obtained. There is no such thing as *the* correlation between mechanical aptitude and abstract intelligence, for instance, but only *a* correlation between certain tests of mechanical aptitude and intelligence, given to certain groups under definite conditions.

* Jones, H. E., *A First Study of Parent-Child Resemblance in Intelligence*, 27th Yearbook of the N.S.S.E. (1928), Part I, 61–72.

† Stead, W. H., and Shartle, C. L., *Occupational Counseling Techniques* (1940), Chapters 7 and 8.

Correlation coefficients are always to be thought of as *conditional* and never as *absolute* indices of relationship.

2. The Interpretation of r in Terms of $\sigma_{(est.)}$ and the Coefficient of Alienation

One of the most practical ways of evaluating the effectiveness of a coefficient of correlation is through the standard error of estimate, $\sigma_{(est.)}$. We have found (p. 320) that $\sigma_{(est. Y)}$ — which equals $\sigma_y\sqrt{1 - r^2}$ — enables us to tell how accurately we can estimate (by means of the regression equation) an individual's score in Test Y when we know his score in Test X. The size of $\sigma_{(est.\ Y)}$ depends directly upon σ_y and upon the correlation between the two tests. When $r = 1.00$, $\sigma_{(est.\ Y)} = .00$, and we can predict a person's score in Y, knowing his score in X, with 100% accuracy — no error. On the other hand, when $r = .00$, $\sigma_{(est.\ Y)} = \sigma_y$, and we can only be certain that the predicted score lies *somewhere* within the limits of the Y-distribution, i.e., within the limits Mean Score $\pm\ 3\sigma_y$. In other words, when $r = .00$ our estimate of a person's Y-score is not aided at all by a knowledge of his score in X. As r decreases from 1.00 to .00, the standard error of estimate increases so markedly that predictions from the regression equation range all the way from certainty to what is virtually a "guess." * The significance of an r, with respect to predictive value, therefore, may be accurately gauged by the extent to which r improves our prediction over a "mere guess."

The following problem will serve as an illustration: Suppose that the correlation between two tests Y and X is .60, and that $\sigma_y = 5.00$. Then $\sigma_{(est.\ Y)}$ is $5 \times \sqrt{1 - .60^2}$ or 4.00. This *SE* is 20% less than 5.00, the $\sigma_{(est.\ Y)}$ when $r = .00$, i.e., when $\sigma_{(est.\ Y)}$ has minimum predictive value. The amount of reduction in

* The term "guess" as here used does not imply an estimate which is based upon no information whatsoever — a shot in the dark, so to speak. When $r = .00$, the most probable Y-score predicted for every individual in the X-distribution is M_Y, and $\sigma_{(est.\ Y)} = \sigma_y$. Hence, our Y-estimates are "guesses" in the sense that they may lie anywhere in the Y-distribution — but not anywhere at all!

$\sigma_{(est. Y)}$ as r varies from .00 to 1.00 is given by the expression $\sqrt{1 - r^2}$, and hence it is possible from $\sqrt{1 - r^2}$ alone to gauge the predictive value of an r. The expression $\sqrt{1 - r^2}$ is often called the *coefficient of alienation* and is denoted by the letter k. The coefficient of alienation may be thought of as measuring the *absence* of relationship between two variables X and Y in the same sense in which r measures the *presence* of relationship. When $k = 1.00$, $r = .00$, and when $k = .00$, $r = 1.00$: the *larger* the coefficient of alienation the *smaller* the degree of relationship, and the less precise the prediction from X to Y. In order to show how the estimate improves as r increases, the k's for certain values of r from .00 to 1.00 are tabulated in Table 51.

TABLE 51

COEFFICIENTS OF ALIENATION (k) FOR VALUES OF r FROM .00 TO 1.00

r	$k = \sqrt{1 - r^2}$	r	$k = \sqrt{1 - r^2}$
.0000	1.0000	.8000	.6000
.1000	.9950	(.8660)	(.5000)
.2000	.9798	.9000	.4359
.3000	.9539	.9500	.3122
.4000	.9165	.9800	.1990
.5000	.8660	.9900	.1411
.6000	.8000	1.0000	.0000
.7000	.7141		
(.7071)	(.7071)		

Note that r must be .866 before k lies *halfway* between 1.00 and .00, before the standard error of estimate is reduced to one-half of its value where $r = .00$. For r's of .80 or less, the coefficients of alienation are clearly so large that predictions of individual scores based upon the regression equation are little better than "guesses." * Even when $r = .99$, the standard error of estimate is still 1/7 as large as when $r = .00$. In contrast to actuarial prediction, therefore, the estimation of an individual's score in one test from another is not warranted unless r is at least .90.

The coefficient E given by the formula below is often useful in

* An r is more efficient in forecasting the probable success of a *group* (see p. 322).

providing a quick estimate of the predictive efficiency of an obtained r. E, which is called the "coefficient of forecasting efficiency" or the coefficient of dependability, is derived from k as follows:

$$E = 1 - \sqrt{1 - r^2} \tag{63}$$

or

$$E = 1 - k$$

("*coefficient of forecasting efficiency*" *or coefficient of dependability*) *

To illustrate the application of E, suppose that the correlation of a test (or of a test battery) with some criterion of performance is .50. From formula (63) $E = 1 - .87$ or .13; and the test's efficiency in predicting criterion scores may be said to be 13%. When $r = .90$, $E = .56$ and the test is 56% efficient; when $r = .98$, $E = .80$ and the test is 80% efficient, and so on. Obviously, the correlation must be above .87 for the test's forecasting efficiency to be greater than 50%.

E gives essentially the same information as $\sigma_{(\text{est. } Y)}$ or k. Thus, if $r = .50$, $k = .87$ and $\sigma_{(\text{est. } Y)}$ is 87% of σ_y, which is its value when $r = .00$. Accordingly, an r of .50 reduces the $\sigma_{(\text{est. } Y)}$ by 13%.

3. The Interpretation of r in Terms of the Coefficient of Determination (r^2)

The interpretation of r in terms of "overlapping" factors in the tests being correlated may be generalized through an analysis of the *variance* (σ^2) of the dependent variable — usually the Y test. In studying the variability among individuals upon a given test, the variance of the test scores is often a more useful measure of "spread" than is the standard deviation.† The object in analyzing the variance of Test Y is to determine

* See Conrad, H. S., and Martin, G. B., "The Index of Forecasting Efficiency, for the Case of a 'True' Criterion," *Journal of Experimental Education*, 4 (1935), 231–244.

† Ezekiel, M., *Methods of Correlation Analysis* (2nd ed., 1941), p. 139 and pp. 211–212.

from the correlation between Y and X what part of Test Y's variance is associated with, or dependent upon, the variance of Test X, and what part is determined by the variance of factors not in Test X.

If we have calculated the correlation between Tests Y and X, σ^2_y gives a measure of the *total* variance of the Y-scores; and $\sigma^2_{(\text{est. } Y)}$, which equals $\sigma^2_y(1 - r^2_{xy})$, gives a measure of the variance *left* in Test Y when that part of the variance produced by Test X is *ruled out* or *held constant.** To illustrate, if we have the correlation between height and weight in a group of school children, $\sigma^2_{(ht)}$ will be reduced to $\sigma^2_{(\text{est. } ht.)}$, when the variance in weight is zero, i.e., when *all* of the children have the same weight. If $\sigma^2_{(\text{est. } Y)}$ is subtracted from σ^2_y there remains that part of the variance of Test Y which *is* associated with Test X; and if this value is divided by σ^2_y, we obtain that fraction of the variance of Test Y *attributable to* or *associated with* Test X. Carrying out the operations described, we have

$$\frac{\sigma^2_y - \sigma^2_{(\text{est. } Y)}}{\sigma^2_y} = \frac{\sigma^2_y - \sigma^2_y + \sigma^2_y r^2_{xy}}{\sigma^2_y} = r^2_{xy}$$

from which it is clear that r^2_{xy} gives the *proportion* of the variance of Test Y which is associated with Test X. When used in this way, r^2 is called the *coefficient of determination.* If the correlation between Tests Y and X is .707, r^2 is .50. Hence, an r of .707 means that 50% of the variance of Test Y is associated with the variability in Test X. Since $r^2 + k^2 = 1.00$,† the proportion of the variance in Test Y which is *not* associated with Test X is given by k^2. In the present case, since r^2 is .50, k^2 is also .50.

The coefficient of determination tells us what part of the variance of Test Y is determined by Test X. But r alone gives us no information as to the character of the association and we cannot assume a causal relationship unless we have evidence beyond the correlation. Inspection of the squares of small coefficients of correlation emphasizes the slight degree of associa-

* See Chapter XIII for further discussion of this topic.

† See Table 51.

tion, in terms of related changes in variability, indicated by low r's. An r of .10, for example, or .20, or .30, between Tests X and Y, indicates that only 1%, 4%, and 9%, respectively, of the variance of Y is associated with X. On the other hand, when r is .95, about 90% ($r^2 = .90$) of the variance of Test Y is associated with Test X, only 10% being unrelated. Valuable insight into the part played by one or more variables in determining the total variance of a criterion may be obtained through the coefficient of determination.

4. Summary

It may be helpful to summarize the main points brought out in this section.

(1) Whether an obtained r is to be regarded as "high," "medium," or "low" will depend upon the variables being studied, the reliability coefficients of the two tests, the size of the group and its variability, and the purpose for which the r is being computed. Correlation coefficients are never *absolute* indices of relationship.

(2) The accuracy with which an r enables us to predict (through the regression equation) *individual* scores in Test Y from given scores in Test X may be determined from $\sigma_{(\text{est. }Y)}$, from E, and from k, *the coefficient of alienation.*

(3) The *coefficient of determination* provides a method of determining what proportion of the total variance (σ^2) of Test Y is associated with Test X; and what proportion is independent of Test X. This method of analysis may be extended to problems employing partial and multiple correlation (p. 425).*

PROBLEMS

1. Write out the regression equations in score form for the correlation table in example 3, page 305.
   (*a*) Compute $\sigma_{(\text{est. }Y)}$ and $\sigma_{(\text{est. }X)}$.
   (*b*) What is the most probable height of a boy who weighs 30

* Wright, Sewall, "Correlation and Causation," *Journal of Agricultural Research,* 20 (1921), 557–585.

pounds? 45 pounds? What is the most probable weight of a boy who is 36 inches tall? 40 inches tall?

2. In example 4, page 305, find the most probable grade made by a child whose score on Army Alpha is 120. What is the $\sigma_{(est.)}$ of this grade?

3. What is the most probable algebra grade of a child whose I.Q. is 100 (data from example 5, p. 306)? What is the $\sigma_{(est.)}$ of this grade?

4. Given the following data for two tests:

History (X)	English (Y)
$M_X = 75.00$	$M_Y = 70.00$
$\sigma_x = 6.00$	$\sigma_y = 8.00$

$$r_{xy} = .72$$

(*a*) Work out the regression equations in score form. ✓

(*b*) Predict the probable grade in English of a student whose history mark is 65. Find the $\sigma_{(est.)}$ of this prediction.

(*c*) If r_{xy} had been .84 (σ's and means remaining the same) how much would $\sigma_{(est.\ Y)}$ be reduced?

5. The correlation of a test battery with worker efficiency in a large factory is .40, and 70% of the workers are regarded as "satisfactory."

(*a*) From seventy-five applicants you select the "best" twenty-five in terms of test score. How many of these should be satisfactory workers?

(*b*) How many of the best ten should be satisfactory?

(*c*) How many in the two groups should be satisfactory if selected at random, i.e., without using the test battery?

6. Plot the regression lines in on the correlation diagram given in example 5, page 306. Calculate the means of the Y-arrays (successive Y-columns), plot as points on the diagram, and join these points with straight lines. Plot, also, the means of the X-arrays and join them with straight lines. Compare these two "lines-through-means" with the two fitted regression lines (see Fig. 52, p. 310).

7. In a group of 115 freshmen, the r between reaction time to light and substitution learning is .30. The σ of the reaction times is 20 ms. What would you estimate the correlation between these

two tests to be in a group in which the σ of the reaction times is 25 ms.?

8. Show the regression effect in example 4, page 305, by calculating the regression equation in standard-score form. For I.Q.'s $\pm$ 1.00σ and $\pm$ 2.00σ from the mean I. Q., find the corresponding school marks in standard-score form.
9. Basing your answer upon your experience and general knowledge of psychology, decide whether the correlation between the following pairs of variables is most probably (1) positive or negative; (2) high, medium, or low.
   (*a*) Intelligence of husbands and wives.
   (*b*) Brain weight and intelligence.
   (*c*) High-school grades in history and physics.
   (*d*) Age and radicalism.
   (*e*) Extroversion and college grades.
10. How much more will an r of .80 reduce a given $\sigma_{(est.)}$ than an r of .40? An r of .90 than an r of .40?
11. (*a*) Determine k and E for the following r's: .35; $-$ .50; .70; .95. Interpret your results.
    (*b*) What is the "forecasting efficiency" of an r of .45? an r of .99?
12. The correlation of a criterion with a test battery is .75. What percent of the variance of the criterion is associated with variability in the battery? What percent is independent of the battery?

Answers

1. $\bar{Y} = .40X + 24.12$; $\bar{X} = 1.26Y - 11.52$
   (*a*) $\sigma_{(est.\ Y)} = 1.78$; $\sigma_{(est.\ X)} = 3.16$
   (*b*) 36.12 inches; 42.12 inches; 33.84 pounds; 38.88 pounds
2. 85.2; $\sigma_{(est.\ Y)} = 7.0$
3. $\bar{X} = .37Y + 8.16$. When Y(I.Q.) is 100, $\bar{X}$ (algebra) is 45.2
   $\sigma_{(est.\ X)} = 6.8$
4. (*a*) $\bar{Y} = .96X - 2$; $\bar{X} = .54Y + 37.2$
   (*b*) 60.4; $\sigma_{(est.\ Y)} = 5.5$
   (*c*) 22%
5. (*a*) 21
   (*b*) 9
   (*c*) 17.5 and 7 (i.e., 70%)
7. $r = .65$
8. $\pm$ .46 and $\pm$ .92

10. Five times as much; seven times as much.

11. (*a*)

r	k	E
.35	.94	.06
$-.50$	.87	.13
.70	.71	.29
.95	.31	.69

(*b*) 11%; 86%

12. 56%; 44%

CHAPTER XI

FURTHER METHODS OF CORRELATION

In Chapters IX and X, we described the linear, or product-moment correlation methods, and showed how, by means of r and the regression equations, one can "predict" or "forecast" values of one variable from a knowledge of the other. The linear correlation coefficient is useful in psychology and education as a measure, primarily, of the relationship between test scores and other determinations of performance. Test scores (as we have seen) represent a series of measurements of a continuous variable taken along a numerical scale. Many situations arise, however, in which the investigator does not have scores and must work with data in which differences in merit or capacity can be expressed only by ranks (e.g., in orders of merit): or by classifying an individual into one of several descriptive categories. This is especially true in vocational and applied psychology and in the field of personality and character measurement. Again, there are problems in which the relationship among the measurements made is *non-linear*, and cannot be described by the product-moment r. In such cases other methods of determining correlation must be employed; and the purpose of this chapter is to develop some of the more useful of these techniques.

I. Computing Correlation from Ranks

Differences among individuals in many traits can often be expressed by *ranking* the subjects in one-two-three order when such differences cannot be measured directly. Persons, for example, may be ranked in order of merit for honesty, athletic ability, salesmanship, or social adjustment when it is impossible to *measure* these complex behaviors. In like manner, various products, or specimens such as advertisements, color combina-

tions, handwriting, compositions, jokes, and pictures which are admittedly hard to measure may be put in order of merit for esthetic quality, beauty, humor, or some other characteristic. In computing the correlation between two series of ranks, special methods which take account of relative position have been devised. These methods may also be applied to *scores* which have been arranged in order of merit. Although our scores represent quantitative determinations on a metric scale, when we have only a few (less than twenty-five for example), it is often advisable to rank them in order of merit and compute the correlation by the rank-difference method instead of by the longer and more laborious product-moment method. Coefficients of correlation calculated from a few cases are not very reliable at best, and their chief value lies in suggesting the possible existence of relationship, as in a preliminary survey. In such situations the rank-difference method will probably give as adequate a result as that obtained by a more refined technique, and is much easier to apply.

1. Calculation of ρ (rho) by the Method of Rank-Difference

The method of rank-difference is illustrated in Table 52. The problem is to find the relationship between the length of service and the selling-efficiency of twelve salesmen. The names of the men (A, B, C, etc.) are listed in column (1) of the table, and in column (2), opposite the name of each man, is given the number of years he has been in the service of the company. In column (3), the men are ranked in order of merit in accordance with their length of service. For example G, who has been longest with the company, is ranked 1; C, whose length of service is next longest, is ranked 2; and so on down the list. Note that both A and J have the same period of service, and that each is ranked 7.5. Instead of ranking the first man 7 and the second man 8, or both 7 or both 8, we compromise by ranking both 7.5 and F, who follows, 9.*

* If three men receive the same rank, e.g., 7, 8, 9, each is ranked 8 and next man in order is ranked 10. If four men receive the same rank, e.g., 7, 8, 9, and 10, each is ranked 8.5 and the next in order 11.

TABLE 52

TO ILLUSTRATE THE RANK-DIFFERENCE METHOD OF MEASURING CORRELATION

(1) Salesmen	(2) Years of Service	(3) Order of Merit (Service)	(4) Order of Merit (Efficiency)	(5) Difference between Ranks (D)	(6) Difference Squared (D^2)
A	5	7.5	6	1.5	2.25
B	2	11.5	12	.5	.25
C	10	2	1	1.0	1.00
D	8	4	9	5.0	25.00
E	6	6	8	2.0	4.00
F	4	9	5	4.0	16.00
G	12	1	2	1.0	1.00
H	2	11.5	10	1.5	2.25
I	7	5	3	2.0	4.00
J	5	7.5	7	.5	.25
K	9	3	4	1.0	1.00
L	3	10	11	1.0	1.00
$N = 12$					58.00

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 58}{12(143)} = .80 \tag{64}$$

In column (4) the men have been ranked by the sales manager in order of merit for efficiency as salesmen: C, the most efficient man, is ranked 1; and B, the least efficient, is ranked 12. In column (5) the difference (designated D) between each man's efficiency rank and his years-of-service rank is entered; and in the last column each of these D's has been squared. Since each D is squared in column (6), no account need be taken of + and − signs in column (5). The correlation between the two orders of merit may now be computed by substituting for ΣD^2 and N in the formula

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} \tag{64}$$

(*rank correlation coefficient*, ρ)

in which D represents the difference in rank of an individual in the two series; ΣD^2 is the sum of the squares of all such differences; and N is the number of cases. Substituting 58 for the ΣD^2 and 12 for N in formula (64), we obtain a ρ of .80. The

symbol ρ (read as *rho*) is the rank order coefficient of correlation. ρ may be transmuted into a product-moment r by means of tables, but the difference between ρ and its equivalent r is so small that with little loss of accuracy ρ may be taken as equal directly to r.

2. The Significance of ρ (rho)

If ρ is small and N reasonably large (thirty or more) the SE of ρ can be determined by the following formula:

$$\sigma_\rho = \frac{1.05(1 - \rho^2)}{\sqrt{N - 1}} \qquad (65)\ *$$

(*standard error of ρ, rank-order coefficient of correlation*)

Whenever N is small, the SE of ρ is likely to be larger than the value given by the formula, as the sampling distribution of ρ is not normal (p. 297). For this reason, a ρ computed from less than thirty cases must always be interpreted with caution. A better method of determining significance, especially when ρ is large, is to test the obtained ρ against the null hypothesis, that is, to use Table 49, page 299. For example, we find that for $N - 2$ or 10 degrees of freedom (Table 49), an r must be .71 to be significant at the .01 level. Since our ρ (or r) of .80 is considerably larger than this value, it is clearly very significant although N is small.

If a calculated ρ is .40, say, and N is 28, the SE_ρ by (65), is .16. As ρ is 2.5 times its SE, from Table 17 it is almost significant at the .01 level and clearly significant at the .05 level. A better test of significance (which does not assume normality of the sampling distribution) is to compute t by formula (53), page 298, viz., $t = \dfrac{.40\sqrt{26}}{\sqrt{1 - .40^2}} = 2.22$. From Table 29, we note that when $N - 2 = 26$, t is 2.06 at $P = .05$ and 2.78 at $P = .01$. Hence, ρ is significant at the .05 level, but is *not* significant at

$$*\ PE_\rho = \frac{.7063(1 - \rho^2)}{\sqrt{N - 1}} \qquad (65a)$$

the .01 level. This same result can be obtained directly from Table 49. We find, for instance, for $N - 2 = 26$ that an r must be .37 to be significant at the .05 level, and .48 to be significant at the .01 level.

3. Summary of the Rank-Difference Method

The product-moment method takes into account the size of the score as well as its position in the series. The rank-difference method takes account only of the positions of the items in the series. No allowance is made for size of gaps between adjacent scores. Individuals, for example, who score 90, 89, and 70 on a given test are ranked 1, 2, 3 in order of merit, although the difference between 90 and 89 is 1, and the difference between 89 and 70 is 19. Considerable accuracy may be lost in translating scores over into ranks, as gaps will appear in the rankings when a number of scores, all of the same size, receive the same rating. The rank-difference method is rarely used with test scores when N is larger than thirty and is often an exploratory device.

II. Measuring Correlation from Data Grouped into Categories

1. Bi-serial Correlation

In many problems it becomes important to calculate the correlation between traits or attributes, when the members of the group can be measured (i.e., given scores) in the first variable, but can only be classified into *two* categories in the second or "dichotomous" variable. (The term dichotomous means "cut into two parts.") We may, for instance, wish to know the correlation between MA and "social adjustment" in a group of nursery-school children, when our subjects have been given scores in the first trait, but are simply classified as "socially adjusted" or "not socially adjusted" in the second trait. Other examples of dichotomous classification with reference to some attribute are athletic-non-athletic, Negro-White, radical-con-

servative, socially minded-mechanically minded, literate-illiterate, above eighth grade in school-below eighth grade, and the like. Many test and questionnaire items also are scored so as to give responses which fall into two categories; as, for example, problems marked Passed or Failed, statements marked True or False, personality inventory items answered Yes or No, interest test items marked Liked or Disliked, and so on. The correlation between a set of scores and a two-category classification (like those listed above) cannot be found by the ordinary product-moment formula or by the rank-difference method. However, if we can assume that the attribute for which we have made a two-way or dichotomous classification would be continuous and normally distributed if more information were available so that classification could be made in finer units, the correlation between such a trait and a set of scores may be computed by the *bi-serial* correlation method.

(1) Calculation of Bi-serial r

The calculation of bi-serial r is illustrated in Table 53. The problem is to find the correlation between total scores on a test and the answers to a single item in the test (Item 72); or put differently, to find whether those who make high scores on the test tend to answer Item 72 "Yes" more often than "No." The first column of Table 53 gives the class-intervals of the score distribution. Column two gives the distribution of scores made by the sixty subjects who answered "Yes" to Item 72, and column three the distribution of scores made by the forty subjects who answered "No." The sum of all of the frequencies on the score-intervals gives the total distribution of 100 cases (column four). The steps in calculating bi-serial r from here on are as follows:

Step 1

Calculate M_p, the mean of the scores made by the sixty subjects who answered "Yes" to Item 72. Also calculate M_q, the mean of the scores made by the forty subjects who answered "No" to Item 72. In our problem, $M_p = 60.08$; and $M_q = 55.00$.

TABLE 53

TO ILLUSTRATE THE CALCULATION OF THE BI-SERIAL r BETWEEN TOTAL SCORES ON A TEST AND THE ANSWERS TO A SINGLE ITEM ON THE TEST

Scores on Test	Responses to Item 72 "Yes"	"No"	f
80–84	3		3
75–79	4	2	6
70–74	6	2	8
65–69	5	5	10
60–64	10	9	19
55–59	10	5	15
50–54	15	5	20
45–49	4	3	7
40–44	3	2	5
35–39		4	4
30–34		2	2
25–29		1	1
	60	40	100
	(p)	(q)	

$M = 58.05$; mean of all scores ($N = 100$)

$\sigma = 11.63$; σ of all scores ($N = 100$)

$M_p = 60.08$; mean of "Yes" responses ($N = 60$)

$M_q = 55.00$; mean of "No" responses ($N = 40$)

$p = .60$; proportion answering "Yes" to Item 72

$q = .40$; proportion answering "No" to Item 72

$z = .386$; height of ordinate separating 60% from 40% in a normal distribution (Table 54).

$$r_{\text{bis}} = \frac{M_p - M_q}{\sigma} \cdot \frac{pq}{z} \qquad (66)$$

$$= \frac{60.08 - 55.00}{11.63} \times \frac{(.60)(.40)}{.386}$$

$$= .27$$

$$\sigma_{r_{\text{bis}}} = \frac{\left(\frac{\sqrt{pq}}{z} - r^2_{\text{bis}}\right)}{\sqrt{N}} \qquad (67)$$

$$= \frac{\left(\frac{\sqrt{.24}}{.386} - (.27)^2\right)}{\sqrt{100}}$$

$$= .12$$

Step 2

Calculate the σ of the whole distribution — the distribution of the 100 scores. This σ, which equals 11.63, gives the spread of the test scores in the entire group.

Step 3

Sixty percent of the group (p) answered "Yes" to Item 72, and 40% (q) answered "No" (p always equals $1 - q$). Assuming a normal distribution of opinion on this item (varying from complete agreement on through indifference to complete disagreement) upon which a dichotomous division has been forced, we place the dividing line between the "Yes" and "No" groups at a distance of 10% from the middle of the curve, as shown in the figure below.

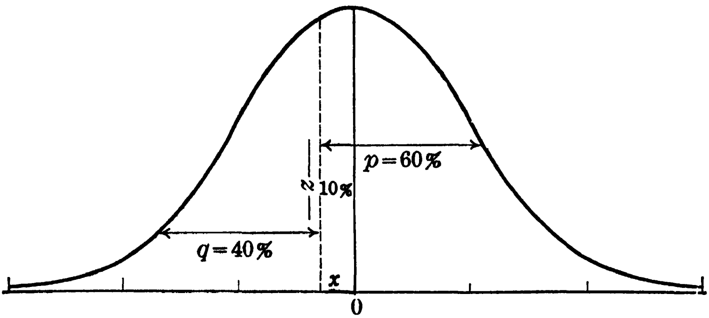


FIG. 55.

From Table 54, the height of the ordinate (i.e., z) which is 10% from the mean of a normal distribution is .386.

Step 4

Having computed M_p, M_q, σ, p, q, and z, we find r_{bis} from the formula

$$r_{\text{bis}} = \frac{M_p - M_q}{\sigma} \times \frac{pq}{z} \tag{66}$$

(*bi-serial coefficient of correlation or bi-serial* r)

in which, as illustrated by the problem above, and shown in Table 53

M_p = mean of the group in the first category (usually the group showing superior or more desirable characteristics)
M_q = mean of the group in the second category
σ = standard deviation of the entire group
p = proportion of the whole group in category one
q = proportion of the whole group in category two ($p = 1 - q$)
z = height of the ordinate in the normal curve dividing p from q

In Table 53, r_{bis} is .27, indicating a tendency, though not a strong one, for "Yes" answers to Item 72 to accompany high total scores.

(2) The *SE* of Bi-serial r

Provided neither p nor q is very small (e.g., smaller than .05), an approximate formula for the standard error of bi-serial r is

$$\sigma_{r_{\text{bis}}} = \frac{\left(\frac{\sqrt{pq}}{z} - r^2_{\text{bis}}\right)}{\sqrt{N}} \tag{67}$$

(*SE of* r_{bis} *for values of* p *and* q *greater than .05*)

The *SE* of the r_{bis} of .27 found in Table 53 is .12, and the critical ratio is .27/.12 or 2.25. From Table 29 we find this bi-serial r to be significant at the .05, but not at the .01 level.*

TABLE 54

DEVIATES (x/σ) IN TERMS OF σ-UNITS AND ORDINATES (z) FOR GIVEN AREAS MEASURED FROM THE MEAN OF A NORMAL DISTRIBUTION WHOSE TOTAL AREA = 1.00

$[x/\sigma = x]$

Area from the Mean (α)	x or (x/σ)	z	Area from the Mean (α)	x or (x/σ)	z
.00	.000	.399	.26	.706	.311
.01	.025	.399	.27	.739	.304
.02	.050	.398	.28	.772	.296
.03	.075	.398	.29	.806	.288
.04	.100	.397	.30	.842	.280
.05	.126	.396	.31	.878	.271
.06	.151	.394	.32	.915	.262
.07	.176	.393	.33	.954	.253
.08	.202	.391	.34	.995	.243
.09	.228	.389	.35	1.036	.233
.10	.253	.386	.36	1.080	.223
.11	.279	.384	.37	1.126	.212
.12	.305	.381	.38	1.175	.200
.13	.332	.378	.39	1.227	.188
.14	.358	.374	.40	1.282	.176
.15	.385	.370	.41	1.341	.162
16	.412	.366	.42	1.405	.149
.17	.440	.362	.43	1.476	.134
.18	.468	.358	.44	1.555	.119
.19	.496	.353	.45	1.645	.103
.20	.524	.348	.46	1.751	.086
.21	.553	.342	.47	1.881	.068
.22	.583	.337	.48	2.054	.048
.23	.613	.331	.49	2.326	.027
.24	.643	.324	.50	∞	.000
.25	.675	.318			

* At the .05 level tne *CR* = 1.98, at the .01 level 2.63, when the (N − 1) = 99

(3) An Alternative Formula for Bi-serial r

There is another — and slightly different — formula for bi-serial r which is often useful. This is

$$r_{\text{bis}} = \frac{M_p - M_T}{\sigma} \times \frac{p}{z} \tag{68}$$

(bi-serial coefficient of correlation or bi-serial r *in terms of* M_T*, the mean of the total group)*

in which

M_p = mean of the group in the first (or p category)
M_T = mean of entire group
σ = standard deviation of entire group
p = proportion of whole group in category one
z = height of ordinate in normal curve dividing p from q

Substituting in formula (68) the values for M_p, M_T, σ, p, and z, shown in Table 53, we have

$$r_{\text{bis}} = \frac{60.08 - 58.05}{11.63} \times \frac{.600}{.386} = .27$$

which checks our previous result.

Formula (68) is especially well-suited to those problems in which sub-groups having different characteristics are drawn from a larger group, the larger group mean (M_T) remaining the same.

The bi-serial correlation method has frequently been used in determining item validity,* that is, in finding whether success or failure upon a given item is correlated with total score in the test or with score in some criterion (Table 53). If those who achieve high scores in the criterion get an item right more often than those who make low scores, the item will be positively correlated with the criterion. Such an item is a better measure of the criterion than one which correlates zero or negatively with criterion scores.

* Long, J. A., and Sandiford, Peter, *The Validation of Test Items,* Department of Educational Research, University of Toronto, Bulletin #3 (1935), 16–17.

When items are scored 1 if correct and 0 if incorrect, the assumption of normality in the distribution of responses to any given item is not warranted.* Formula (69) below gives a bi-serial coefficient which does not assume continuity in the distribution of single test items, and is recommended for use in item analysis:

$$r_{\text{bis}} = \frac{M_p - M_q}{\sigma} \cdot \sqrt{pq} \tag{69}$$

(bi-serial coefficient of correlation for use in item analysis)

Formula (68) may be — and is generally — used in determining item validity, but (69) is somewhat more defensible mathematically, and is easier to apply. The validity-index of Item 72 (Table 53) by formula (69) is .21.

2. Tetrachoric Correlation

We have seen in the last section that when one variable is continuous and is expressed in the form of test scores, and the other is dichotomous or in a twofold classification, bi-serial r gives a measure of the relationship between the two variables. An extension of this problem to which bi-serial r is not applicable presents itself when *both* variables are dichotomous. We then have a 2×2 or fourfold table, from which a modified form of the product-moment coefficient, called *tetrachoric r*, may be calculated. Tetrachoric r is useful when one wishes to find the relationship between two characters or attributes neither of which is directly measurable, but both of which are capable of being separated into two categories. Thus, if we wish to measure the correlation between school attendance and employment, persons might be classified into those who have attended high school and those who have not; and into those who are employed and those who are unemployed. Or if we wish to discover the correlation between intelligence and social maturity, children might be classified as "above average" and "below average" in

* Richardson, M. W., and Stalnaker, J. L., "A Note on the Use of Bi-serial r in Test Research," *Journal of General Psychology,* 8 (1933), 463–465.

intelligence, on the one hand, and as socially mature and socially immature on the other. The tetrachoric correlation method assumes that the two variables being studied are essentially *continuous*, and would be *normally distributed* if it were possible to classify them more exactly into finer groupings.

(1) Calculation of Tetrachoric r

Table 55 illustrates a 2×2 fold table, and shows the steps involved in calculating tetrachoric r. The problem is to find whether a larger number of successful than of unsuccessful salesmen tend to be "socially well adjusted." The data are artificial. The X-variable (along the top of the diagram) is divided into two categories "successful" and "unsuccessful"; and the Y-variable (along the left of the diagram) is divided into two categories "socially well adjusted" and "socially poorly adjusted." The sums of the rows show that sixty salesmen $(a + b)$ out of the sample of 100 are classed as well adjusted socially, and that forty salesmen $(c + d)$ are classed as poorly adjusted socially.* The proportions in each category (p and q) are 60% and 40%, respectively. The sums of the columns show that fifty-five of the 100 salesmen are classified as unsuccessful, and forty-five as successful; the proportions are 55% (q') and 45% (p'). On the assumption that "social adjustment" is distributed normally, from the proportions $p = .60$, and $q = .40$, we obtain an $x = -.253$, and $z = .386$. These last two values are read from Table 54 as follows: The perpendicular line (i.e., the ordinate, z) separating the *upper* 60% from the *lower* 40% in a normal curve is just 10% from the mean. Hence, entering the first column of Table 54 with $\alpha = .10$, we read $x = -.253$, and $z = .386$. See diagram on page 356.

The x' and z' values corresponding to $p' = .45$ and $q' = .55$ are calculated in the same way. The perpendicular line dividing

* To accord with the plan of the ordinary correlation table (p. 280), the categories in Table 55 have been so arranged that concentration of data in the *first* and *third* quadrants (a and d) denotes positive correlation; concentration of data in the *second* and *fourth* (b and c) quadrants negative correlation.

TABLE 55

To Illustrate the Calculation of Tetrachoric r (r_t)
(The data are hypothetical)

X-variable

Y-variable		100 Salesmen		Totals
		Unsuccessful	Successful	
	Socially Well Adjusted	25 (*b*)	35 (*a*)	60 $p = 60\%$
	Socially Poorly Adjusted	30 (*d*)	10 (*c*)	40 $q = 40\%$
	Totals	55 $q' = 55\%$	45 $p' = 45\%$	100

For $p = .60$, $q = .40$, $\alpha = .10$
$x = -.253$ [Table 54]
$z = .386$ [Figure 56]

For $p' = .45$, $q' = .55$, $\alpha = .05$
$x' = .126$ [Table 54]
$z' = .396$ [Figure 56]

$$\frac{ad - bc}{N^2zz'} = r + \frac{xx'r^2}{2}$$

$$\frac{1050 - 250}{100^2(.386)(.396)} = r + \frac{(-.253)(.126)r^2}{2}$$

$$.523 = r - .016r^2$$

or

$$.016r^2 - r + .523 = 0^*$$

$$r = \frac{+1 \pm \sqrt{1 - 4(.016)(.523)}}{2 \times .016} = \frac{+1 \pm \sqrt{1 - .033472}}{.032}$$

$$= \frac{+1 \pm .9831}{.032}$$

$= .53$ (taking numerator as $+1 - .9831$)
$= +62$ (taking numerator as $+1 + .9831$)

* The general form of a quadratic equation is $ax^2 + bx + c = 0$. The two values of x (i.e., the roots of the equation) may be computed by the formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

In the equation $.016r^2 - r + .523 = 0$, $a = .016$; $b = -1.00$; and $c = .523$. Hence,

$$r = \frac{+1 \pm \sqrt{1 - 4(.016)(.523)}}{2 \times .016}$$

$= .53$ or 62 (an impossible value)

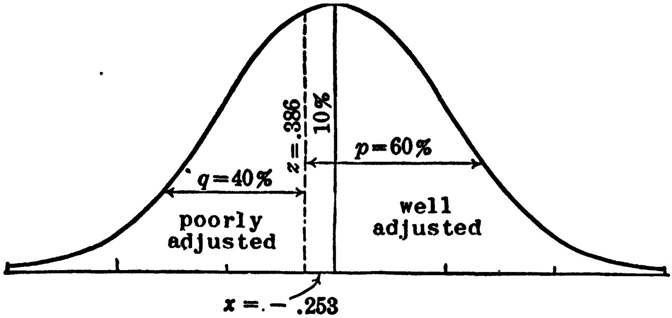


FIG. 56.

the upper 45% (the percent successful) from the lower 55% (the percent unsuccessful) is 5% from the mean; and from Table 54, for $\alpha = .05$, $x' = .126$, and $z' = .396$. See diagram below.

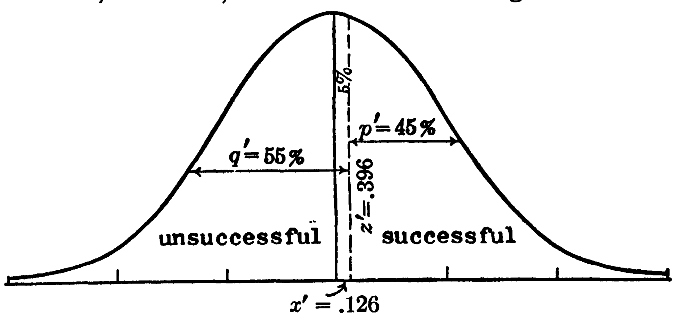


FIG. 57.

An approximate formula for tetrachoric r may be written as follows:

$$\frac{ad - bc}{N^2 zz'} = r_t + \frac{xx'r^2_t}{2} \tag{69a}$$

(*approximate formula for tetrachoric* r) *

in which

x and x' = σ-distances from the means to the points separating the proportion in the upper category from the proportion in the lower category;

* Pearson, Karl, *On the Correlation of Characters Not Quantitatively Measurable*, Philosophical Transactions, Royal Society of London, Series A, 195 (1900), 1–47.

z and z' = the heights of the ordinates *at* the points of division;
a, b, c, d = entries in the four cells, see Table 55;
N = number of cases; i.e., sum of entries in the four cells.
r_t = the tetrachoric coefficient of correlation.

In Table 55, *ad* is found to equal 1050, and *bc* to equal 250. Substituting for these quantities, and for x, x', z, z' and N^2 in formula (69a), we obtain $r_t = .53$. This coefficient indicates a fairly substantial correlation between success in salesmanship and social adjustment. In order to compute r_t it is necessary that we solve a quadratic equation. The method of carrying through this solution is given in Table 55 and in the footnote at the bottom of the table. Note that only the first of the two solutions for r_t is a possible value, as the second is greater than unity.

The investigator who finds it necessary to calculate many tetrachoric r's may greatly shorten his work by using the computing diagrams devised by Thurstone and his co-workers.* These charts enable one to obtain a solution for r_t by graphic methods as soon as the proportion within each of the four cells of the table is known.

(2) The *SE* of a Tetrachoric r

The formula for σ_{r_t} is an exceedingly complex expression and is not reproduced here. The derivation of σ_{r_t} will be found in books dealing with the mathematics of statistical theory.† The computation of σ_{r_t} can be greatly shortened by the use of Pearson's Tables XXIII and XXIV.‡ An approximation to the *SE* of a tetrachoric r may be found in the following way: The σ_{r_t} is about 50% higher than the *SE* of an equivalent product-moment r, that is, a product-moment r equal to the given r_t and calculated from a sample of the same size as that upon which

* Chesire, L., Saffir, M., Thurstone, L. L., *Computing Diagrams for the Tetrachoric Correlation Coefficient*, University of Chicago Bookstore (1933).

† Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases* (1940), pp. 370–375.

‡ Pearson, Karl, *Tables for Statisticians and Biometricians* (1914), Introduction, xl–xli, and p. 35.

r_t is based. The SE of a product-moment r of .53 is .07, for $N = 100$; hence the SE of a tetrachoric r of .53 is approximately $.07 \times 1.5$ or .11. The obtained r_t of .53 is nearly five times its SE and is, therefore, significant at the .01 level (Table 17).

Tetrachoric r is often used as a means of evaluating a test's efficiency in separating two contrasted or "criterion" groups. An example is given in Table 56 (the data are artificial). The problem is to find whether a test of deductive reasoning (here, a

TABLE 56

To Illustrate the Use of Tetrachoric r in Evaluating a Given Test

$N = 125$

X-variable

Y-variable	College Juniors: Non-Science Majors	College Juniors: Science Majors	
Above Test Mean	24% (b)	35% (a)	$p = 59\%$
Below Test Mean	29% (d)	12% (c)	$q = 41\%$
	$q' = 53\%$	$p' = 47\%$	100%

For $p = .59$, $q = .41$
$x = -.228$
$z = .389$

For $p' = .47$, $q' = .53$
$x' = .075$
$z' = .398$

$$\frac{.1015 - .0288}{(.389)(.398)} = r + \frac{(-.228)(.075)r^2}{2} \qquad (69a)$$

$$.470 = r - .009r^2$$

or

$$.009r^2 - r + .470 = 0$$

$$r = \frac{+1 \pm \sqrt{1 - 4(.009)(.470)}}{2(.009)}$$

$$= \frac{+1 \pm .9915}{.018}$$

$$= .47, \text{ or } 111 \text{ (an impossible value)}$$

syllogism test) will differentiate fifty-nine college juniors majoring in science from sixty-six college juniors majoring in literature

or languages (non-science). The X-variable is divided into science majors and non-science majors; the Y-variable into those above and those below the mean of the test, i.e., the mean score established by the entire junior class. The entries in the cells, a, b, c, and d, are expressed in percents, so that N^2 in formula (69a) is 1.00. As shown in Table 56, the correlation between majoring in science and high scores on the syllogism test is .47. If one were investigating a number of tests with a view toward determining their relative values as indicators of scientific aptitude, the worth of each test could be measured in accordance with its ability to separate the two criterion groups.*

3. The Contingency Coefficient

The coefficient of contingency, C, is often used to determine relationship when the variables under study can be put into more than two classes or categories. The contingency coefficient may be derived directly from χ^2 (p. 241); but it differs from χ^2 in that it provides a measure of *correlation* which under certain conditions (p. 363) is comparable to the product-moment r. C bears the following relation to χ^2:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \tag{70}$$

(a formula for C*, the contingency coefficient, in terms of* χ^2*)*

Calculating χ^2 and applying formula (70), we find that the C of example 6 (1) (p. 376) is $\sqrt{\dfrac{31.43}{513 + 31.43}}$ or .24. Taken at face value, this C indicates a small degree of relationship between marriage-adjustment and education of husbands. To find whether the obtained C is indicative of a significant relationship, we should calculate its standard error. Unfortunately, the SE of C is a complex expression † and is somewhat laborious

* For a discussion of the application of tetrachoric r to problems involving two widely separated or extreme groups in which the middle group is eliminated, see Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases* (1940), pp. 375–384.

† See Kelley, T. L., *Statistical Method* (1923), p. 269.

to compute. For a $C = .00$, however, $\sigma_C = \frac{1}{\sqrt{N}}$; and this formula may be employed to give a rough test of the significance of an obtained C. On the null hypothesis the relationship between marriage-adjustment and education is .00, and its SE is $\frac{1}{\sqrt{513}}$ or .044. Our calculated C of .24 is $\frac{.24}{.044}$ or nearly $6SE_C$ removed from a C of .00. Hence, $C = .24$ may be considered to indicate a small but highly significant degree of correlation between marriage-adjustment and education of husbands.

When one is not directly interested in χ^2 itself, it is possible to compute C directly rather than by way of χ^2. There are two methods of calculating C which will be given in order.

(1) Method A of Calculating C

Table 57 illustrates the computation of C from a 4×4 fold contingency table. The table gives the classification of 1000 fathers and sons with respect to eye color. The independence values for each cell have been computed as shown in Table 40 (p. 252). To repeat the method of calculation, 335/1000 of *all* sons are described as blue-eyed. This proportion of 358 (i.e., $\left.\frac{335 \times 358}{1000}\right)$ gives 120 as the number of fathers who might be expected to have blue-eyed sons "by chance," as contrasted with the 194 fathers who actually did have blue-eyed sons. When the independence values have been found, we square each obtained cell entry, and divide by its own independence value as shown in Table 57. The sum of these quotients gives S; and from S and N, C is calculated by the formula

$$C = \sqrt{\frac{S - N}{S}} \tag{71}$$

(formula for C, *coefficient of contingency, calculated directly)*

In Table 57, C is .46. On the null hypothesis when $C = .00$, $\sigma_C = \frac{1}{\sqrt{1000}}$ or .03. The obtained C of .46 is fifteen times σ_C

TABLE 57

To Illustrate the Calculation of C, the Coefficient of Contingency by Method A

Father's Eye Color

Son's Eye Color	Blue	Gray	Hazel	Brown	Totals
Blue	(120) 194	(88) 70	(60) 41	(66) 30	335
Gray	(102) 83	(75) 124	(51) 41	(56) 36	284
Hazel	(49) 25	(36) 34	(25) 55	(27) 23	137
Brown	(87) 56	(64) 36	(44) 43	(48) 109	244
Totals	358	264	180	198	1000

I. Independence Values

$$\frac{335 \times 358}{1000} = 120 \qquad \frac{137 \times 358}{1000} = 49$$

$$\frac{335 \times 264}{1000} = 88 \qquad \frac{137 \times 264}{1000} = 36$$

$$\frac{335 \times 180}{1000} = 60 \qquad \frac{137 \times 180}{1000} = 25$$

$$\frac{335 \times 198}{1000} = 66 \qquad \frac{137 \times 198}{1000} = 27$$

$$\frac{284 \times 358}{1000} = 102 \qquad \frac{244 \times 358}{1000} = 87$$

$$\frac{284 \times 264}{1000} = 75 \qquad \frac{244 \times 264}{1000} = 64$$

$$\frac{284 \times 180}{1000} = 51 \qquad \frac{244 \times 180}{1000} = 44$$

$$\frac{284 \times 198}{1000} = 56 \qquad \frac{244 \times 198}{1000} = 48$$

II. Calculation of C

$$\frac{(194)^2}{120} = 313.6$$
$$\frac{(83)^2}{102} = 67.5$$
$$\frac{(25)^2}{49} = 12.8$$
$$\frac{(56)^2}{87} = 36.0$$
$$\frac{(70)^2}{88} = 55.7$$
$$\frac{(124)^2}{75} = 205.0$$
$$\frac{(34)^2}{36} = 32.1$$
$$\frac{(36)^2}{64} = 20.3$$
$$\frac{(41)^2}{60} = 28.0$$
$$\frac{(41)^2}{51} = 33.0$$
$$\frac{(55)^2}{25} = 121.0$$
$$\frac{(43)^2}{44} = 42.0$$
$$\frac{(30)^2}{66} = 13.6$$
$$\frac{(36)^2}{56} = 23.1$$
$$\frac{(23)^2}{27} = 19.6$$
$$\frac{(109)^2}{48} = 247.5$$

$$S = 1270.8$$
$$N = 1000$$
$$S - N = 270.8$$

$$C = \sqrt{\frac{S - N}{S}} = \sqrt{\frac{270.8}{1270.8}} = .46$$

and hence is highly significant of a fairly strong correlation between eye color in father and son.

C may be either plus or minus, the sign to be affixed depending upon an inspection of the contingency table itself. In Table 57 it is evident that pigmentation of eyes in father and son is positively correlated * and hence that C is positive.

A disadvantage of the contingency coefficient is the fact that C does not remain constant for the same data when the number of classes varies. The C calculated from a 3×3 fold table will not ordinarily equal the C calculated from the same data arranged in, say, a 5×5 fold table. Moreover, the maximum value which C can take will depend upon the fineness of the classification employed. It can be shown † that

when the number of classes = 2, C cannot exceed .707
when the number of classes = 3, C cannot exceed .816
when the number of classes = 4, C cannot exceed .866
when the number of classes = 5, C cannot exceed .894
when the number of classes = 6, C cannot exceed .913
when the number of classes = 7, C cannot exceed .926
when the number of classes = 8, C cannot exceed .935
when the number of classes = 9, C cannot exceed .943
when the number of classes = 10, C cannot exceed .949

In the light of this table, Yule suggests that we "restrict the use of the 'coefficient of contingency' to 5×5 fold or finer classifications" in order that the maximum value of C may be as near unity as possible. At the same time, we should avoid a too-fine classification or C will be affected by slight or "casual irregularities of no physical significance"; and, in addition, the arithmetic of calculation will be greatly (and needlessly) in-

* We note, for example, that 194 blue-eyed fathers have blue-eyed sons, while only 30 brown-eyed fathers have blue-eyed sons. Moreover, 109 brown-eyed fathers have brown-eyed sons while only 56 blue-eyed fathers have brown-eyed sons. Comparisons of this sort will show that association between pigmentation in the eyes of father and son is positive.

† Yule, G. U., and Kendall, M. G., *An Introduction to the Theory of Statistics* (12 ed., 1940), p. 69.

creased. Pearson * has worked out a correction for "broad categories" which should be applied to C's calculated from 4×4 fold or broader groupings if C is to be compared with r. For 5×5 fold or finer classifications, this correction is so small that for practical purposes it may be disregarded.

Since the classification in Table 57 is 4×4 fold, the value of C will be increased if corrected for broad categories. An approximate correction, which is easier to apply than Pearson's correction, can be made by dividing the obtained C by the maximum value which C can take in a 4×4 fold contingency table. In the present problem, dividing our C of .46 by .866 (the maximum C for a 4×4 fold table) we obtain a "corrected C" of .53. This value may be taken as approximately equal to r; it indicates a fairly high correlation between pigmentation of eyes in father and son.

The relation of C to r is, under certain conditions, very close. C is substantially equivalent to r (1) when the grouping is relatively fine — 5×5 fold or finer; (2) when the sample is large; (3) when the two variables may legitimately be classified into categories; and (4) when we are justified in assuming that the variables under investigation are normally distributed.

(2) Method B for Calculating C

The arithmetic involved in computing C may be lessened somewhat by combining the twofold process of (1) calculating independence values and (2) dividing the square of each cell frequency by its independence value. This method is illustrated in Table 58. The first occupied cell in the first column of the table has a frequency of 1 and an independence value of $\frac{99 \times 8}{384}$; hence the cell frequency squared and divided by the independence value is $\frac{1 \times 384}{8 \times 99}$. This fraction, namely, $\frac{1 \times 384}{8 \times 99}$, is the

* Pearson, Karl, "On the Measurement of the Influences of 'Broad Categories' on Correlation," *Biometrika*, 9 (1913), 130; also see the discussion in Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases* (1940), pp. 391–393.

TABLE 58

TO ILLUSTRATE THE CALCULATION OF C BY METHOD B
BOYS: AGES 4–5 YEARS

Weight in Pounds

Height in Inches	24–28	29–33	34–38	39–43	44–48	49–53	Total
45–47			1		2		3
42–44			4	35	21	5	65
39–41		5	87	90	7	1	190
36–38	1	18	72	8			99
33–35	5	15	5				25
30–32	2						2
	8	38	169	133	30	6	384

Column 1: $\frac{1}{8}\left[\frac{1}{99}+\frac{25}{25}+\frac{4}{2}\right] = .3762$

Column 2: $\frac{1}{38}\left[\frac{25}{190}+\frac{324}{99}+\frac{225}{25}\right] = .3264$

Column 3: $\frac{1}{169}\left[\frac{1}{3}+\frac{16}{65}+\frac{7569}{190}+\frac{5184}{99}+\frac{25}{25}\right] = .5549$

Column 4: $\frac{1}{133}\left[\frac{1225}{65}+\frac{8100}{190}+\frac{64}{99}\right] = .4671$

Column 5: $\frac{1}{30}\left[\frac{4}{3}+\frac{441}{65}+\frac{49}{190}\right] = .2792$

Column 6: $\frac{1}{6}\left[\frac{25}{65}+\frac{1}{190}\right] = .0650$

$P = \overline{2.0688}$

$$C = \sqrt{\frac{P-1}{P}} = \sqrt{\frac{1.0688}{2.0688}} = .72$$

contribution of this particular cell to the total S. In the same way, the contribution to S of the next cell in this column is found to be $\frac{5^2 \times 384}{8 \times 25}$, and of the third and last cell, $\frac{2^2 \times 384}{8 \times 2}$. These contributions from column 1 may be combined to give $\frac{384}{8}\left(\frac{1}{99}+\frac{25}{25}+\frac{4}{2}\right)$. The contribution of each of the other five

columns to S may be found in like manner. Moreover, since N (i.e., 384) is a common factor in each column, it may be left out of the computations entirely, in calculating the contribution of each cell, as shown in Table 58. Then if the sum of all six columns is denoted by P,

$$C = \sqrt{\frac{P-1}{P}} \tag{72}$$

(*alternate method of calculating* C)

In Table 58, C equals .72 and the coefficient of correlation, r, from the same table is .71 (see p. 305). The correspondence of C and r here is very close, closer perhaps than that generally to be expected, although the difference between the two coefficients is never very great when the conditions prescribed on page 363 are met. In the present case, N is large, the classification is 6×6 fold, and the distributions are fairly normal.

III. Curvilinear or Non-Linear Relationship

1. The Correlation Ratio

The relationship between the paired values of two sets of measures, X and Y, may be described in a general way as "linear" or "non-linear." When the means of the arrays of the successive columns and rows in a correlation table follow straight lines (at least approximately), the regression is said to be linear or straight-line (p. 281). When the drift or trend of the means of the arrays (columns or rows) cannot be well described by a straight line, but can be represented by a *curve* of some kind, the regression is said to be curvilinear or in general non-linear.

Our discussion in Chapter IX was concerned entirely with linear relationship, the extent or degree of which is measured by the product-moment coefficient of correlation, r. It sometimes happens in mental measurement, however, that the relationship between two variables is definitely non-linear; and when this is true, r is not an adequate measure of the degree of correspondence or correlation. When the regression is non-linear, a curve

joining the means of successive arrays (in the columns, say) will fit these mean values more exactly than will a straight line. Hence, should a truly curvilinear relationship be described by a straight line, the scatter or spread of the paired values about the regression line will be greater than the scatter about the better-fitting regression curve. The smaller the spread of the paired scores about the regression line or the regression curve which relates the variables X and Y (or Y and X), the higher the relationship between the two variables. For this reason, an r calculated from a correlation table in which the regression is curvilinear will *always be less* than the true relationship. An example will make this situation clearer. The correlation between the following two short series, as given by the product-

Variable X	Variable Y
1	.25
2	.50
3	1.00
4	2.00
5	4.00

moment formula, is $r = .93$ [formula (46), p. 289]. The *true* correlation between the two series, however, is clearly perfect, since changes in Y are directly related to changes in X. As X increases by 1 (i.e., in arithmetic progression) Y doubles (i.e., increases in geometric progression). The reason why r is less than 1.00 becomes obvious as soon as we plot the paired X and Y values. As shown in Figure 58, the relationship between X and Y is curvilinear, and is exactly described by a curve which passes through the successively plotted points. When linear relationship is forced upon these data, the plotted points do not fall along the straight line, and the product-moment coefficient, r, is less than 1.00. However, the correlation-ratio, or coefficient of non-linear relationship η (read as *eta*) for the given data is 1.00.

Eta measures the concentration of paired X- and Y-values about a relation *curve* just as r measures the concentration of

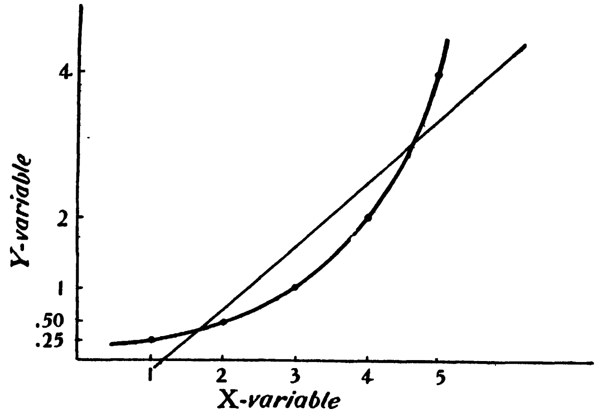


FIG. 58. To Illustrate Non-Linear Relationship.

paired values about a relation *line*. *Eta* is a more general coefficient than r as it is applicable when regression is linear as well as when it is non-linear. If the regression is linear and the means fall along straight lines, η will equal r. But if regression is non-linear and the means lie along a curve, η will be greater than r. The coefficient of correlation, therefore, is a limiting value of the more general coefficient, η, just as straight-line relationship is a limiting case of curvilinear relationship. There are always two η's in every non-linear correlation table, just as there are always two regression coefficients, $r\,\frac{\sigma_y}{\sigma_x}$ and $r\,\frac{\sigma_x}{\sigma_y}$, in a table in which regression is linear. The first correlation-ratio, written η_{yx}, gives the regression of Y on X (Y is the dependent variable). The second correlation-ratio, written η_{xy}, gives the regression of X on Y (X is the dependent variable). (Compare with the two regression equations (p. 310) in a correlation table in which relationship is linear.)

The correlation-ratio is always given the positive sign, and its value lies between .00 and 1.00. Whether the *direction* of relationship given by η is positive, negative, or a varying one, must be determined by inspection of the correlation diagram.

2. The Calculation of η in a Correlation Table

One of the most useful methods of calculating the two η's (η_{xy} and η_{yx}) in a correlation table in which the relationship is known (or suspected) to be non-linear is illustrated in Figure 59.* Ordinarily, one will wish to compare the two calculated η's with the r obtained from the same data in order to determine whether regression is, or is not, significantly non-linear. For this reason, the computation of r is included in Figure 59 as part of the process of calculating the η's. The steps to be followed in finding η_{yx} may be outlined briefly. The method of calculating η_{xy}, shown on the right of the diagram, follows exactly the method outlined here for the calculation of η_{yx} and hence will not be repeated.

Step 1

Construct a correlation table as shown in Figure 59, and described on page 276. Calculate σ_y and σ_x using the Assumed Mean method (p. 41).

Step 2

Determine the entries in the $\Sigma y'$ row. These entries are found, as described on page 286, by multiplying the frequency in each column by its deviation (i.e., its y') measured *in units of class-interval* from the Assumed Mean of the Y-distribution. To illustrate, in column one, reading down, we have $(1 \times -2) + (1 \times -4) + (4 \times -5) + (2 \times -6)$ or -38. For column two, the $\Sigma y'$ entry is $(1 \times 2) + (2 \times -1) + (2 \times -2) + (2 \times -3) + (1 \times -4)$ or -14. Square each $(\Sigma y')$ entry to give the $(\Sigma y')^2$ row. Then divide each entry in the $(\Sigma y')^2$ row by its corresponding f_x to give the row $\frac{(\Sigma y')^2}{f_x}$. In column one, for example, divide 1444 by 8 to obtain 180.50; and in column two, divide

* For further discussion of the method here outlined, see Dvorak, A., "A Simplified Computation of Non-linear Correlation," *Journal of Educational Research*, 25 (1932), 99–104.

Holzinger, K. J., "A Combination Form for Calculating the Correlation Coefficient and Ratios," *Journal of the American Statistical Association*, 18 (1923), 623–627.

Test Scores	Age: 9 to 9-11	10 to 10-11	11 to 11-11	12 to 12-11	13 to 13-11	14 to 14-11	15 to 15-11	16 to 16-11	f_y	y'	fy'	fy'^2	$\Sigma x'$	$(\Sigma x')^2$	$\frac{(\Sigma x')^2}{f_y}$	$\Sigma x'y'$
50-54					1				1	3	3	9	0	0	0	0
45-49		1	1	3	3	2	2		12	2	24	48	−2	4	.33	−4
40-44			3	4	6	4	2		19	1	19	19	−2	4	.21	−2
35-39			3	4	7	5	4		23	0	0	0	3	9	.39	0
30-34		2	4	1	2	4	5		18	−1	−18	18	−1	1	.06	1
25-29	1	2	2				3	1	9	−2	−18	36	−5	25	2.78	10
20-24		2		1		1	2	2	8	−3	−24	72	4	16	2.00	−12
15-19	1	1						1	3	−4	−12	48	−4	16	5.33	16
10-14	4							1	5	−5	−25	125	−13	169	33.80	65
5-9	2								2	−6	−12	72	−8	64	32.00	48
f_x	8	8	13	13	19	16	18	5	100		−63	447	−28		76.90	122
x'	−4	−3	−2	−1	0	1	2	3								
fx'	−32	−24	−26	−13	0	16	36	15	= −28							
fx'^2	128	72	52	13	0	16	72	45	= 398							
$\Sigma y'$	−38	−14	−3	6	13	1	−11	−17	= −63							
$(\Sigma y')^2$	1444	196	9	36	169	1	121	289								
$(\Sigma y')^2/f_x$	180.50	24.50	.69	2.77	8.89	.06	6.72	57.80	= 281.93							
$\Sigma x'y'$	152	42	6	−6	0	1	−22	−51	= 122							

check

$$c_y = -\frac{63}{100} = -.63; \quad c^2_y = .40$$

$$c_x = -\frac{28}{100} = -.28; \quad c^2_x = .08$$

$$\sigma_y = \sqrt{\frac{447}{100} - .40} \times 5 = 2.02 \times 5 \qquad \sigma_x = \sqrt{\frac{398}{100} - .08} \times 1 = 1.975 \times 1$$

$$\eta_{yx} = \frac{\sqrt{\frac{281.93}{100} - .40}}{2.02} = .77 \qquad \eta_{xy} = \frac{\sqrt{\frac{76.90}{100} - .08}}{1.975} = .42$$

$$r = \frac{\frac{122}{100} - (-.28)(-.63)}{2.02 \times 1.975} = .26$$

FIG. 59. To Illustrate Non-Linear Regression and the Calculation of η (*eta*)
(Correlation between age and score on an achievement test)

196 by 8 to obtain 24.50. The total of the $\frac{(\Sigma y')^2}{f_x}$ row in Figure 59 is 281.93.

Step 3

From $\frac{(\Sigma y')^2}{f_x}$, c^2_y, N, and σ_y, calculate η_{yx} by the following formula*:

$$\eta_{yx} = \frac{\sqrt{\dfrac{\Sigma\left[\dfrac{(\Sigma y')^2}{f_x}\right]}{N} - c^2_y}}{\sigma_y} \tag{73}$$

(*correlation-ratio,* η_{yx}, *a measure of non-linear relationship in terms of the standard deviation of the means of the* Y*-arrays*)

In Figure 59 $\frac{(\Sigma y')^2}{f_x} = 281.93$; $N = 100$; $c^2_y = .40$; and $\sigma_y = 2.02$ (in units of interval). Substituting these values in formula (73) we obtain .77 as the value of η_{yx}.

The formula for η_{xy}, the second *eta* in a correlation table, is

$$\eta_{xy} = \frac{\sqrt{\dfrac{\Sigma\left[\dfrac{(\Sigma x')^2}{f_y}\right]}{N} - c^2_x}}{\sigma_x} \tag{74}$$

(*correlation-ratio,* η_{xy}, *a measure of non-linear relationship in terms of the standard deviation of the means of the* X*-arrays*)

In the present problem, $\eta_{xy} = .42$ (see Fig. 59 for calculations).

In most correlation tables, as illustrated here, the two η's will differ in size, since their values depend upon the scatter about the curve joining the means of the Y-arrays, and the scatter about the curve joining the means of the X-arrays. In any particular problem, also, one correlation-ratio will ordinarily be of greater interest than the other; just as in linear cor-

* There are several alternate formulas, equivalent to formula (73), which may be used in calculating η. See Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases* (1940), pp. 312–330; also Yule, G. U., and Kendall, M. G., *An Introduction to the Theory of Statistics* (12 ed., 1940), pp. 242–246.

relation, one regression equation is usually of greater interest than the other (p. 318). In Figure 59, η_{yx} is obviously more valuable than η_{xy}, since it gives the *change* in score (Y) resulting from *changes* in age (X); Y is the dependent variable, and X is the independent variable (p. 314). The curve which describes the relation between age and score — the curve through the means of the Y-columns — has been sketched in on the correlation diagram. Note that this curve begins and ends low, reaching its peak in the middle of the age range. Both younger and older children in the grade make low scores, the highest scores being achieved by children in the middle of the age range. A probable reason for the obtained non-linear relationship between age and score is that the given test contains elements unfamiliar to, or inadequately learned by, the younger children, and items too difficult for the older (and probably duller) children. The best scores, therefore, are achieved by those in the middle of the age range. The product-moment r in Figure 59 is .26.

3. The Standard Error of η

The SE of a correlation-ratio may be calculated by the formula

$$SE_{\eta} = \frac{(1 - \eta^2)}{\sqrt{N - 1}} \tag{75}$$

(*standard error of a correlation-ratio*, η)

The SE of the η_{xy} of .42 is .08, and of the η_{yx} of .77 is .04. Both of these coefficients are clearly significant (see p. 297).

4. The Correction of an Obtained η

The size of an obtained η depends directly upon the fineness of grouping in the X- and Y-variables, as well as upon the size of N. When N is comparatively small and the number of arrays in X or Y is large, a correction* should be applied to the obtained η. The formula for a corrected η is

* Pearson, Karl, "On the Correction Necessary for the Correlation-Ratio η," *Biometrika*, 14 (1923), 412–417.

See also, Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases* (1940), pp. 312–325.

$$\text{Corrected } \eta = \sqrt{\frac{\eta^2_{obt} - \dfrac{(\kappa - 3)}{N}}{1 - \dfrac{(\kappa - 3)}{N}}} \tag{76}$$

(*correction of η for fineness of grouping*)

in which κ equals the number of arrays in X or Y. To illustrate, if we apply this correction to η_{yx} obtained from Figure 59, we have, upon substituting .77 for η_{yx}, 8 for the number of Y-arrays (i.e., columns), and 100 for N,

$$\text{Corrected } \eta_{yx} = \sqrt{\frac{(.77)^2 - .05}{1 - .05}}$$
$$= .76$$

The correction here is small, since η_{yx} is large, and the number of Y-arrays moderate. The correction which must be applied to η_{xy} is larger. Thus, substituting .42 for η_{xy}, and 10 for the number of X-arrays (i.e. rows), we have

$$\text{Corrected } \eta_{xy} = \sqrt{\frac{(.42)^2 - .07}{1 - .07}}$$
$$= .34$$

When η is small, and the grouping fine (i.e., data classified into many intervals) the correction given by formula (76) may be considerable, and hence should be made.

5. Test for Linearity of Regression

It is not always easy to tell from the appearance of a correlation table whether regression is linear or non-linear. In Figure 59, from the curve joining the means of the columns, it seems clear that the regression of Y on X, at least, is non-linear. Further evidence of non-linearity is offered by the fact that the coefficient of correlation calculated from Figure 59 is .26, very much smaller than the η_{yx} of .77. As stated on page 367, when regression is strictly linear $\eta = r$; and the greater the departure of the regression from linearity, in general the greater the discrepancy between η and r.

A test for non-linearity of regression in terms of χ^2 enables us to estimate the significance of the departure of curvilinear relationship from linear relationship. The formula for χ^2 is

$$\chi^2 = (N - \kappa)\left(\frac{\eta^2 - r^2}{1 - \eta^2}\right) \tag{77}$$

(χ^2-*test for linearity of regression*)

in which N is size of sample and κ is number of columns or rows.

In Figure 59, $\eta_{yx} = .77$, $r = .26$, and $\kappa = 8$ (number of columns or Y-arrays). Hence from (77)

$$\begin{aligned}\chi^2 &= (100 - 8)\left(\frac{.59 - .07}{1 - .59}\right)\\ &= 116.8\end{aligned}$$

Entering Table 32 with $\kappa - 2$ or 6 degrees of freedom, we obtain a P which is much smaller than .01. The probability is quite remote, therefore, that a deviation of η_{yx} from r as large as that obtained (i.e., 116.8) could have arisen from sampling accidents. Hence we must abandon the hypothesis of linear relationship and accept the regression as curvilinear.

In Table 59, the second *eta*, η_{xy}, is .42 and $\kappa = 10$ (number of rows or X-arrays). From (77) we find

$$\begin{aligned}\chi^2 &= (100 - 10)\left(\frac{.18 - .07}{1 - .18}\right)\\ &= 11.7\end{aligned}$$

and entering Table 32 with $\kappa - 2 = 8$, we get a P which lies between .30 and .20. The χ^2 of 11.7, therefore, is *not* significant and the regression of X on Y is probably rectilinear — or at least not markedly non-linear.

6. Summary on η and r.

True non-linear relationship is often encountered in psychophysics and in experiments dealing with fatigue, practice, forgetting, and learning. Whenever an experiment is carried on to the point of diminishing returns, relationship will be curvilinear. Most mental and educational tests, when administered to large samples, exhibit linear or approximately linear relationships; and for this reason, r has been employed in psychology and

education to a far greater extent than has η. If regression is significantly non-linear it makes considerable difference whether η or r is the measure of relation. But if the correlation is low and the regression not significantly curvilinear, r will give about as adequate a measure of relationship as η.

The coefficient of correlation has the advantage over η in that knowing r we can write down at once the straight-line regression equation connecting X and Y or Y and X. This is not possible with the correlation ratio. In order to estimate one variable from another (say, Y from X) when regression is non-linear, a curve must be fitted to the means of the Y-columns. The equation of this curve then serves as a "regression equation" from which estimates can be made.*

PROBLEMS

1. Compute the correlation between the following two series of test scores by the rank-difference method and test its significance.

Individual	Intelligence Score (Army Alpha)	Cancellation Score (A-Test + Number Group Checking Test)
1	185	110
2	203	98
3	188	118
4	195	104
5	176	112
6	174	124
7	158	119
8	197	95
9	176	94
10	138	97
11	126	110
12	160	94
13	151	126
14	185	120
15	185	118

[Note: The cancellation scores are in *seconds;* hence the two smallest scores numerically (i.e., 94) are highest and are ranked 1.5 each.]

* Snedecor, G. W., *Statistical Methods* (1940), Chapter 14.

2. Check the product-moment correlations obtained in problems 6 and 7, pages 306–307, Chap. IX, by the rank-difference method.
3. The following data give the distributions of scores on the Thorndike Intelligence Examination made by entering college freshmen who presented 12 or more recommended units, and entering freshmen who presented less than 12 recommended units. Compute bi-serial r by formula (66) and test its significance.

Thorndike Scores	12 or more recommended units	Less than 12 recommended units
90–99	6	0
80–89	19	3
70–79	31	5
60–69	58	17
50–59	40	30
40–49	18	14
30–39	9	7
20–29	5	4
	186	80

4. The following data give the distributions of scores on Army Alpha made by those who answered 50% or more, and those who answered less than 50% of the items in test 2 ("Arithmetic") correctly. Compute bi-serial r and test its significance.

Army Alpha Scores	Subjects answering 50% or more of the items on test 2 correctly	Subjects answering less than 50% of the items on test 2 correctly
185–194	7	0
175–184	16	0
165–174	10	6
155–164	35	15
145–154	24	40
135–144	15	26
125–134	10	13
115–124	3	5
105–114	0	5
	120	110

5. Compute the tetrachoric r's for the following tables which show the
   (1) Relation of alcoholism and health in 811 fathers and sons. Entries are expressed as proportions.
   (2) Correspondence of Yes and No answers to two items of a neurotic inventory.

(1)

		Sons		
		Unhealthy	Healthy	Totals
Fathers	Non-Alcoholic	.343	.405	.748
	Alcoholic	.102	.151	.252
	Totals	.445	.556	1.000

(2)

		Question 1		
		No	Yes	Totals
Question 2	Yes	83	187	270
	No	102	93	195
	Totals	185	280	465

6. Calculate the coefficient of contingency, C, for each of the three tables given below.

(1)

		Marriage-Adjustment Score of Husbands				
		Very Low	Low	High	Very High	Totals
Education of Husbands	Graduate work	4	9	38	54	105
	College	20	31	55	99	205
	High School	23	37	41	51	152
	Grade School	11	10	11	19	51
	Totals	58	87	145	223	513

(2)

		Kind of Music Preferred					
		English	French	German	Italian	Spanish	Totals
Nationality of Subject	English	32	16	75	47	30	200
	French	10	67	42	41	40	200
	German	12	23	107	36	22	200
	Italian	16	20	44	76	44	200
	Spanish	8	53	30	43	66	200
	Totals	78	179	298	243	202	1000

(3)

Education	Salary 0–900	901–1200	1201–2000	2001–4000	4001–10,000	10,001–	Totals
Post Graduate Work				4	1		5
College Graduate			1	30	5	1	37
Business College		1	15	6	1		23
High School	2	10	30	7	1		50
Junior High	7	42	27	3	1		80
Elementary School	19	48	4	1			72
Totals	28	101	77	51	9	1	267

7. The following table shows the relationship between scores upon the Thorndike Intelligence Examination and certain extra-curricular activities of 102 Columbia College students.
   (*a*) Compute η_{yx} and η_{xy}, and the *SE*'s.
   (*b*) Find corrected values for both η's.
   (*c*) Test both η's for linearity of regression.

Thorndike Scores (X)

Extra-curricular activi (X) ties	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–94	95–99	100–104	
18–20					2	2					4
15–17				2		3	1				6
12–14				4		6	2		2		14
9–11		1	2		4	4	6	7	3		27
6–8	1			6	2	2	6	2	4	1	24
3–5	1		1	3	5	3		5	1	1	20
0–2		1		1			1	1	1	2	7
	2	2	3	16	13	20	16	15	11	4	102

8. In the following table (*a*) calculate the two η's and (*b*) test for linearity of regression.

Age in Months (X)

Scores on Test Y	80–89	90–99	100–109	110–119	120–129	130–139	140–149	150–159	f_y
75–79								10	10
70–74								12	12
65–69								18	18
60–64							8	16	24
55–59							10	8	18
50–54							12		12
45–49							14		14
40–44							6		6
35–39						8	6		14
30–34						19	7		26
25–29				2	2	22	5		31
20–24			1	10	17	26			54
15–19		2	4	8	15	12			41
10–14	5	5	12	8	24	9			63
5–9	9	8	16	16	9	9			67
0–4	6	6	3	20	13	7			55
f_x	20	21	36	64	80	112	68	64	465

Answers

1. $\rho = .19$. Not significant (Table 49)
3. $r_{\text{bis}} = .34$. $SE_{r_{\text{bis}}} = .07$; very significant. $P < .01$
4. $r_{\text{bis}} = .47$. $SE_{r_{\text{bis}}} = .07$; very significant. $P < .01$
5. (1) $r_t = -.09$
   (2) $r_t = .33$

6. (1) $C = .24$
   (2) $C = .40$
   (3) $C = .70$

7. (a) $\eta_{yx} = .43$, $SE_{\eta_{yx}} = .08$ (b) η_{yx} (corrected) $= .35$
   $\eta_{xy} = .20$, $SE_{\eta_{xy}} = .10$ η_{xy} (corrected) $= .00$
   (c) $r = -.09$. For η_{yx}, χ^2 by (77) is 19.96. $P < .01$; departure from linearity significant. For η_{xy}, $\chi^2 = 3.14$. P lies between .70 and .50; departure from linearity not significant.

8. (a) $\eta_{yx} = .93$, $SE = .007$ η_{yx} (corrected) $= .93$
   $\eta_{xy} = .82$, $SE = .016$ η_{xy} (corrected) $= .81$
   (b) $r = .78$. For η_{yx}, $\chi^2 = 849.1$. $P < .01$; departure from linearity very significant. For η_{xy}, $\chi^2 = 81.72$. $P < .01$; departure from linearity very significant.

CHAPTER XII

THE RELIABILITY AND VALIDITY OF TEST SCORES

I. THE RELIABILITY OF TEST SCORES

THE reliability of a test, as of any measuring instrument, depends upon the *consistency* with which it gauges the abilities of those to whom it has been applied. When a test is reliable, scores made by the members of a group — upon retest with the same test or with alternate forms of the same test — will differ very little or not at all from their original values. A reliable test is relatively free of chance errors of measurement, and scores earned on it are stable and trustworthy. If a subject scores 84, say, on a reliable test, we feel confident that this score represents very closely his true ability. Scores made on an unreliable test, on the other hand, are subject to large errors of measurement and are neither stable nor trustworthy. When a test is unreliable, subsequent testings will reveal many discrepancies between scores achieved by the same persons on different occasions.

1. Methods of Determining Test Reliability

There are three procedures in common use for determining the reliability (sometimes called the self-correlation) of a test. These are (1) the test-retest (repetition) method; (2) the alternate or parallel forms method; and (3) the split-half method. In addition to these three, a fourth method — the method of "rational equivalence" — is also being widely used. All of these procedures furnish "estimates" of the reliability of test scores; sometimes one method and sometimes another will give the best estimate.

(1) Test-Retest (Repetition) Method

Repetition of a test is the simplest method of determining reliability: the test is given and then repeated on the same group and the correlation is calculated between the first and second sets of scores. While the test-retest method is sometimes the only feasible procedure, it is open to various objections. If the test is repeated immediately, many subjects will recall their first answers and spend their time on new material, thus increasing their scores. Besides the memory effect, practice and the confidence induced by familiarity with the material will almost certainly affect scores when one takes a test for the second time. Transfer effects are likely to be different from person to person. If the net effect of transfer is to make for closer agreement between scores achieved on the first and second giving of a test than would otherwise be the case, the reliability coefficient will be too high. When a sufficient time interval has elapsed between the first and second administrations of the test to offset (in part, at least) memory, practice, and other effects, the reliability coefficient will be a closer estimate of the actual consistency of test scores. If the interval between tests is long, however (say, six months or so), and the subjects are children, growth or maturity changes will affect the retest.

The test-retest method will estimate less accurately the reliability of tests which contain novel features and which are highly susceptible to practice than it will the reliability of tests involving routine operations little affected by practice. Because of the difficulty in controlling the conditions which influence scores on different administrations of a test, the test-retest method is used less generally than are the other two methods.

(2) Alternate or Parallel Forms Method

When alternate or parallel forms of a test have been constructed, the correlation between Form A, say, and Form B is taken as a measure of the self-correlation of the test. This method is employed by the authors of most standard pyscho-

logical and educational tests, for which alternate forms are usually available.

The alternate forms method is usually satisfactory if sufficient time intervenes between the administration of the two forms to weaken or eliminate memory and practice effects. When Form B of a test follows Form A very closely, scores on the second test will usually be increased through practice and familiarity. When such increases are approximately constant (say, three to five points for each score) the reliability coefficient of the test will not be affected, since paired A and B scores maintain their same relative positions in the two distributions. When the mean increase due to practice has been determined, a constant amount can be subtracted from Form B scores to make them comparable to Form A scores.* In drawing up alternate forms of a test, one should be careful to match test materials for content, difficulty, and form; but one must be careful not to make the test forms too much alike. If alternate forms are practically identical, the reliability coefficient of the test will be too high; while if parallel forms are not sufficiently "duplicate" the reliability coefficient will be too low.

(3) The Split-half Method

In the split-half method the test is broken into two equivalent parts and the correlation of these half-tests is computed. From the half-test reliability, the self-correlation of the whole test is estimated by the Spearman-Brown formula described on page 388.

The split-half method is employed when it is not feasible to construct an alternate form of the test nor wise to repeat the test. This situation occurs with many performance tests, as well as with tests and questionnaires dealing with personality traits, attitudes, and the like. A performance test (e.g., picture

* In the Otis Self-Administering Test of Mental Abilities, Higher Examination, for instance, the author suggests that when Form B, which is slightly more difficult than Form A, is given first, four points be added to each score. This is to make scores equivalent to the norms for Form B when this test is given after Form A, as it usually is. See *Manual of Directions*, Otis S-A Test (1928), p. 2.

completion, puzzle solving, form board) is often a very different task when repeated, as the child is familiar with procedure and content. Likewise, many personality tests cannot be given in alternate form nor repeated because of radical changes in the subject's attitude and interests when taking such tests for the second time.

The split-half method is generally regarded as the best of the methods for determining test reliability. Perhaps its main advantage is that all of the data for determining test reliability are obtained upon *one* occasion; hence variations introduced by differences between the two testing situations are eliminated. A disadvantage to the split-half method is that chance errors may affect the scores on both halves of the test in the same way, thus tending to make the reliability coefficient too high. The longer the test, the less the probability that the effects of temporary and variable disturbances will be cumulative and in one direction, and the more accurate the estimate of reliability.

Objection has been raised to the split-half method on the ground that a test can be divided into two parts in a variety of ways so that the reliability coefficient is not a unique value. This criticism is strictly true only when items are of equal difficulty. When items are in strict order of merit from least to most difficult, the split into odds and evens gives a unique determination of the reliability coefficient.

(4) The Method of "Rational Equivalence"

The method of rational equivalence* represents an attempt to get an estimate of the reliability of a test, free from the objections raised against the methods outlined above. Two forms of a test are defined as "equivalent" when corresponding items *a*, *A*, *b*, *B*, etc., are interchangeable; and when the inter-item correlations are the same for both forms. The method of rational

* Kuder, G. F., and Richardson, M. W., "The Theory of the Estimation of Test Reliability," *Psychometrika*, 2 (1937), 151–160.

Richardson, M. W., and Kuder, G. F., "The Calculation of Test Reliability Coefficients Based upon the Method of Rational Equivalence," *Journal of Educational Psychology*, 30 (1939), 681–687.

equivalence stresses the intercorrelations of the items in the test and the correlations of the items with the test as a whole. Four formulas for determining test reliability have been derived, of which the one given below is perhaps the most useful:

$$r_{11} = \frac{n}{(n-1)} \times \frac{\sigma^2_t - \Sigma pq}{\sigma^2_t} \tag{78}$$

(reliability coefficient of a test in terms of the difficulty and the intercorrelations of test items)

in which:

r_{11} = reliability coefficient of whole test;
n = number of items in the test;
σ_t = the SD of the test scores;
p = the proportion of the group answering a test item correctly;
$q = (1 - p)$ = the proportion of the group answering a test item incorrectly.

To apply formula (78) the following steps are necessary:

Step 1

Compute the *SD* of the test scores for the whole group, namely, σ_t.

Step 2

Find the proportions passing *each* item (p) and the proportions failing *each* item (q).

Step 3

Multiply p and q for each item and sum for all items. This gives Σpq.

Step 4

Substitute the calculated values in formula (78).

To illustrate, suppose that a test of sixty items has been administered to a group of eighty-five subjects; $\sigma_t = 8.50$ and $\Sigma pq = 12.43$. Applying (78) we have

$$r_{11} = \frac{60}{59} \times \frac{72.25 - 12.43}{72.25} = .842$$

which is the reliability coefficient of the test.

A simple approximation to formula (78) has been devised.* This formula is useful to teachers and others who want to determine quickly the reliability of short objective classroom examinations or other tests. It reads:

$$r_{11} = \frac{n\sigma^2_t - M(n - M)}{\sigma^2_t\,(n - 1)} \tag{79}$$

[*approximation to formula (78)*]

in which

r_{11} = reliability of the whole test;
n = number of items in the test;
σ_t = *SD* of the test scores;
M = the mean of the test scores.

Formula (79) is a labor saver since only the mean, *SD* and number of items in the test need be known in order to get an estimate of reliability. The correlation need not be computed between alternate forms or between halves of the test. Suppose that an objective test of forty multiple-choice items has been administered to a small class of students. An item answered correctly is scored 1, an item answered incorrectly is scored 0. The mean test score is 25.70 and $\sigma_t = 6.00$. What is the reliability coefficient of the test? Substituting in (79), we have

$$r_{11} = \frac{40 \times 36.00 - 25.70(40 - 25.70)}{36.00 \times 39}$$
$$= .76$$

The assumption is made in formula (79) that all test items have the same degree of difficulty, i.e., that the same *proportion* of subjects (but not necessarily the same *persons*) pass each item. In a power test items are never of equal difficulty. Formula (79) will give a satisfactory approximation to the test's reli-

* Froelich, G. J., "A Simple Index of Test Reliability," *Journal of Educational Psychology*, 32 (1941), 381–385.

ability, however, even when the test items cover a wide range of difficulty. Formula (79) always underestimates to a slight degree the reliability of a test as found by the split-half technique and the Spearman-Brown formula, and the more widely items vary in difficulty the greater the underestimation. This formula provides a minimum estimate of reliability — we may feel sure that the test is at least as reliable as we have found it to be by (79).

Formulas (78) and (79) are not strictly comparable to the three methods for determining the reliability of test scores given above. In a sense, these formulas provide an estimate of the internal consistency of the test rather than an estimate of the dependability of test scores. The method of rational equivalence is superior to the split-half technique in certain theoretical aspects, but differences in reliability as found by the two methods are never very large (of the order .02, etc.) Formula (79) is often to be preferred to the split-half method because of the time and calculation it saves rather than for other reasons.

2. Factors Influencing the Reliability of Test Scores: Chance and Constant Errors

Many factors affect the reliability of a test besides fluctuations in interest and attention, shifts in emotional attitude, and the differential effects of memory and practice. To these "psychological" factors must be added environmental disturbances such as distractions, noises, interruptions, errors in scoring, and the like. All of these variable influences (environmental and psychological) are subsumed under the head "chance errors." Errors, to be truly "chance," must influence a score in such a way as to cause it to vary above — as often as below — its "true" value. The reliability coefficient is a quantitative estimate of the importance of chance or variable influences upon test scores.

Constant errors, as distinguished from chance errors, work in only one direction. Constant errors may raise or lower all of the scores on a retest or on the alternate forms of the test,

but will not affect the reliability coefficient. If every paper on Form B of a test is scored 5 points too high, for example, the self-correlation of the test will not be affected (i.e., the correlation between Form A and Form B) but all of the scores on the second form will be in error by 5 points.

How high should the self-correlation of a test be in order for the reliability of the test to be considered satisfactory? This is an important question, and its answer depends upon the nature of the test, the size and variability of the group tested, and the purpose for which the test was given. To distinguish reliably between the means of two relatively small groups of narrow range of ability (for example, a fifth grade and a sixth grade) a reliability coefficient need be no higher than .50 or .60. If the test is to be used to differentiate among the individuals in the group, however, its reliability should be .90 or more. Most of the authors of intelligence tests and educational achievement examinations report correlations of .90 or more between alternate forms of their tests. Since the self-correlation of a test is directly affected by the variability within the group, in reporting a test's reliability coefficient the standard deviation of the group should always be given.

3. The Effect upon Reliability of Lengthening or Repeating a Test

(1) The Reliability of Coefficient from Many Applications or Repetitions of a Given Test

The mean of five determinations of height will, in general, be more reliable than a single determination (p. 183), and the mean of ten determinations will (in general) be more reliable than the mean of five. On the same principle, increasing the length of the test, or averaging the results obtained from several applications of the test, or from alternate forms, will tend to increase reliability. If the self-correlation of a test is not satisfactory what will be the effect of doubling or tripling the test's length? To answer this question experimentally would require considerable time and labor. Fortunately, a good measure of

the effect of lengthening or repeating a test may be obtained from the Spearman-Brown "prophecy formula":

$$r_{nn} = \frac{nr_{1\text{I}}}{1 + (n - 1)\, r_{1\text{I}}} \tag{80}$$

(*Spearman-Brown formula for estimating the correlation between* n *forms of a test, and* n *other similar forms*)

in which

r_{nn} = the correlation between n forms of a test and n alternate forms (or the mean of n forms against the mean of n other forms);

$r_{1\text{I}}$ = the reliability coefficient.

The subscripts ("1I") show that the correlation is between two forms of the *same* test.

To illustrate the use of formula (80) suppose that in a group of 100 adults the self-correlation of a test is .70. What will be the effect upon test reliability of tripling the length of the test? Substituting $r_{1\text{I}} = .70$ and $n = 3$ in formula (80) and solving for r_{nn}, we have

$$r_{nn} = \frac{3 \times .70}{1 + 2 \times .70} = \frac{2.10}{2.40} = .88$$

Tripling the test's length, therefore, increases its reliability coefficient from .70 to .88. Instead of tripling the length of the test we could give *three* parallel forms of the test and average the three scores made by each person. The reliability of these mean scores (each based upon three measures) will be the same, as far as purely statistical factors are concerned, as the reliability got by tripling the length of the test.

The prophecy formula may also be used to find how many times a test should be repeated in order for test scores to reach a given standard of reliability. Suppose that the self-correlation of a test is .80. How much will the test have to be lengthened, or how many times repeated, in order to insure a reliability coefficient of .95? Substituting $r_{1\text{I}} = .80$ and $r_{nn} = .95$ in the formula, and solving for n, we have

$$.95 = \frac{.80n}{1 + .80n - .80} = \frac{.80n}{.20 + .80n}$$

and

$$n = 4.75 \text{ or } 5 \text{ in whole numbers}$$

The test must be five times its present length, therefore, or *five* alternate forms must be given and averaged, before the self-correlation of the test will reach .95.

Predictions of test reliability by the Spearman-Brown formula are valid only when the items or questions added to the test cover the same ground, are of equal range of difficulty, and are comparable in other respects to the items of the original test. When these conditions are satisfied, there would appear to be no reason, as far as the mathematical process is concerned, why we could not boost the self-correlation of a test to any desired figure, simply by continuing to increase its length or by continuing to repeat it. But it is highly improbable that the reliability coefficient of a test could be so increased indefinitely. In the first place, it is impracticable if not impossible to increase a test's length, say, ten or fifteen times. Furthermore, beyond a certain point, boredom, fatigue, loss of incentive, and the like inevitably affect our results and lead to "diminishing returns." When the material added to the test is strictly comparable to the original test items, and when motivation remains substantially constant, the experimental evidence* indicates that a test may be increased to six or seven times its original length, and the Spearman-Brown formula will still give a close estimate of empirically determined results. But after the first four or five lengthenings the prophecy formula may "over-predict" — give higher estimated reliabilities than those obtained by actual calculation. This is not an especially serious drawback, however, as a test which needs so much lengthening in order to yield

* Holzinger, K. J., and Clayton, B., "Further Experiments in the Application of Spearman's Prophecy Formula," *Journal of Educational Psychology*, 16 (1925), 289–299.

Ruch G. M., Ackerson, Luton, and Jackson, J. D., "An Empirical Study of the Spearman-Brown Formula as Applied to Educational Test Material," *Journal of Educational Psychology*, 17 (1926), 309–313.

reliable results should be radically changed in form or content, or better still, perhaps, discarded in favor of another test.

The Spearman-Brown formula may be applied to ratings, judgments, and other estimates as well as to test items. When measuring the reliability of a personality rating scale, for instance, by correlating the ratings made by two equally competent judges, we may employ the prophecy formula to estimate the increased reliability which might be expected if there were four, six or more judges.*

(2) The Reliability Coefficient from One Application of a Test

When a test has no alternate form and cannot well be repeated, we may calculate the reliability of *half* of the test and then proceed to estimate the reliability of the *whole* test by the Spearman-Brown formula. This method is called the "split-half technique" (p. 382). The procedure is to make up two sets of scores by combining, say, alternate exercises or items in the test. The first set of scores represents, for example, performance on the odd-numbered items, 1, 3, 5, 7, etc.; and the second set of scores performance on the even-numbered items, 2, 4, 6, 8, etc. Other ways of making the two halves of the test as comparable as possible in content, difficulty, and susceptibility to practice may be employed, but the method described is the one most commonly used. From the self-correlation of the half test, the reliability coefficient of the whole test may be estimated from the formula

$$r_{1\mathrm{I}} = \frac{2r_{\frac{1}{2}\frac{\mathrm{I}}{\mathrm{II}}}}{1 + r_{\frac{1}{2}\frac{\mathrm{I}}{\mathrm{II}}}} \tag{81}$$

(Spearman-Brown formula for estimating reliability from two comparable halves of a test)

in which

$r_{1\mathrm{I}}$ = the reliability coefficient of the whole test;

$r_{\frac{1}{2}\frac{\mathrm{I}}{\mathrm{II}}}$ = the reliability coefficient of one-half of the test, found experimentally.

* Clark, E. L., "Spearman-Brown Formula Applied to Ratings of Personality Traits," *Journal of Educational Psychology*, 26 (1935), 552–555.
Remmers, H. H., Shock, N. W., and Kelly, E. L., "An Empirical Study of the Validity of the Spearman-Brown Formula as Applied to the Purdue Rating Scale," *Journal of Educational Psychology*, 18 (1927), 187–195.

When the reliability coefficient of one-half of a test ($r_{\frac{1}{2}\frac{1}{\text{II}}}$) is .60 it follows from formula (81) that the reliability of the whole test ($r_{1\text{I}}$) is .75.

4. The Index of Reliability

An individual's "true score" on a test (p. 181) is defined as the mean of a very large number of determinations made of the given person on the same test or parallel forms of the test administered under approximately identical conditions. The correlation between a series of obtained scores and their corresponding theoretically "true" scores may be found by the formula

$$r_{1\infty} = \sqrt{r_{1\text{I}}} \tag{82}$$

(correlation between obtained scores on a given test and true scores in the function measured by the test)

in which

$r_{1\text{I}}$ = the reliability coefficient of the given test;
$r_{1\infty}$ = the correlation between obtained and true scores.

The symbol "∞" (infinity) designates "true scores," that is, scores obtained from an "infinite" number of administrations of the test to the same group.

The coefficient $r_{1\infty}$ is called the *index of reliability;* it measures the trustworthiness of test scores by showing how well obtained scores agree with their theoretically true counterparts. The index of reliability gives the maximum correlation which the given test is capable of yielding. This follows from the fact that "the highest possible correlation which can be obtained (except as chance might occasionally lead to higher spurious correlation) between a test and a second measure is with that which truly represents what the test actually measures, that is, the correlation between the test and the true scores of individuals in just such tests."*

To illustrate the application of the index of reliability, suppose that for a given test the self-correlation is .64. Then

* Kelley, T. L., "The Reliability of Test Scores," *Journal of Educational Research*, 3 (1921), 327.

$r_{1\infty} = \sqrt{.64}$ or .80; and .80 is the highest correlation of which this test is capable, since it represents the relationship between obtained test scores and true test scores in the same function. If the self-correlation of a test is only .25, so that $r_{1\infty} = \sqrt{.25}$ or .50, it is obviously a waste of time to continue using this test without lengthening or otherwise improving it. A test whose index of reliability is only .50 is an extremely poor estimate of the function which it is trying to measure.

5. The Standard Error of an Obtained Score

The effects of variable or chance errors in producing divergencies of obtained scores from their true counterparts may be estimated by the formula

$$\sigma_{1\infty} = \sigma_1\sqrt{1 - r_{11}} \tag{83}$$

(*standard error of an obtained score*)

in which

$\sigma_{1\infty}$ = the standard error of an obtained score (sometimes called the "standard error of measurement");
σ_1 = the standard deviation of the test scores;
r_{11} = the reliability coefficient of the test.

The subscript "$_{1\infty}$" indicates this standard deviation to be a measure of the error made in taking an obtained score (i.e., 1) as an estimate of the true score (i.e., ∞). To illustrate the use of $\sigma_{1\infty}$ suppose that in a group of 300 college freshmen the reliability coefficient of an aptitude test in mathematics is .92 and the *SD* of this distribution is 15.00. From formula (83) we have

$$\sigma_{1\infty} = 15\sqrt{1 - .92} = 4.2 \text{ or 4 in whole numbers}$$

and the odds are 2:1 that the obtained score made by *any* individual in the group does not differ from his true score by more than $\pm$ 4 points. If subject *AB* has a score of 85, we may feel confident (the chances are .95) that his score "actually" lies between 77 and 93 ($\pm$ 1.96 $\times$ 4.2).* Generalizing for the entire group, we should expect about two-thirds of the 300 scores

* See page 185.

to be in error by 4 points or less; the other one-third (or 100) to be in error by more than 4 points.

The reader should note carefully the difference between $\sigma_{(est.)}$ (see p. 320) and $\sigma_{1\infty}$. The first formula enables us to say with what degree of assurance we can predict an individual's score on *one* test when we know his score on a *second* (and usually a different) test. The actual prediction of the most probable score is made, of course, by way of the regression equation connecting the two variables (p. 317). The *SE* of an obtained score, $\sigma_{1\infty}$, is also an estimate formula; it tells us how adequately an obtained score represents the true score. Although the true score is unknown, we can, nevertheless, tell from $\sigma_{1\infty}$ how much our obtained score probably misses the true value. The *SE* of an obtained score is the best method of expressing the reliability of a test, since it takes account of the *self-correlation* of the test as well as of the *variability* within the group.

Formula (83) provides a general estimate of the *SE* of any score over the entire range of the test. When the range is wide, the agreement of scores on two forms of the test may differ considerably at successive parts of the scale. To refine our estimate of the reliability of our test scores, we may compute $\sigma_{1\infty}$ for different levels of achievement. This has been done for the new Stanford-Binet; the $\sigma_{1\infty}$ for I.Q.'s 130 and above, for example, is 5.24, for I.Q.'s 90–109, 4.51, for I.Q.'s 70 and below, 2.21, etc. The method is described in the references given below.*

6. The Dependence of the Reliability Coefficient upon the Size and Variability of the Group

The reliability coefficient of a test administered to a small group (a single grade, say), cannot be compared directly with the reliability coefficient of the same test administered to a larger group, e.g., to the children in several grades. The self-correla-

* Terman, L. M., and Merrill, M. A., *Measuring Intelligence* (1937), p. 46.

McNemar, Quinn, "The Expected Average Difference between Individuals Paired at Random," *Journal of Genetic Psychology*, 43 (1933), 438–439.

tion of a test (like any correlation coefficient) is affected by the variability of the group; and the larger and more heterogeneous the group, the greater test variability tends to be. If we know the self-correlation of a test in a narrow range (ordinarily a small group) we can estimate the self-correlation of the same test in an increased range (ordinarily a larger group) by the formula

$$\frac{\sigma_s}{\sigma_l} = \frac{\sqrt{1 - r_{ll}}}{\sqrt{1 - r_{ss}}} \tag{84}$$

(relation between σ's and reliability coefficients obtained in different ranges when the test is equally effective throughout both ranges)

in which

σ_s and σ_l = the σ's of the test scores in the small and large groups, respectively;

r_{ss} and r_{ll} = the reliability coefficients in the small and large groups.

To illustrate the use of formula (84) suppose that for a single fifth grade, $r_{ss} = .50$, and $\sigma_s = 5.00$; and that for a larger group made up of children from grades three to seven, $\sigma_l = 15.00$. Assuming our test to be as effective in the large group as in the small, what is the reliability coefficient of the test in the large group? If we substitute for σ_s, σ_l and r_{ss} in formula (84), $r_{ll} = .94$. This means that a reliability coefficient of .50 in the small group indicates as high a degree of test consistency as a reliability coefficient of .94 in a group in which the score range is three times as wide.

II. The Validity of Test Scores

The validity of a test, or of any measuring instrument, depends upon the *fidelity* with which it measures whatever it purports to measure. A homemade yardstick is valid when measurements made by it are proved to be accurate by standard measuring rods. And in the same way a test is valid when the capacity which it gauges corresponds to the same capacity as

otherwise objectively measured and defined. The difference between validity and reliability can be made clear, perhaps, by an illustration. Suppose a clock is set forward twenty minutes. If the clock is a good timepiece, the time it "tells" will be reliable (i.e., consistent), but it will not be valid as judged by "standard time." The reliability of the measurements made by scales, thermometers, yardsticks, chronoscopes, clocks, etc., is determined by making repeated measurements of the same facts; and validity is determined by comparing the measures returned by the given instrument with highly precise (if arbitrary) "standard" measures. The reliability of mental measures is found in the same way. But since precise and independent "standards" (criteria) are rarely found in mental measurement, the validity of a test can never be estimated as precisely as can the validity of a thermometer or a rheostat.

1. The Determination of Validity through Correlation with a Criterion

The validity of a test is determined directly, whenever possible, by finding the correlation between the test and some independent criterion. A criterion is an objective measure in terms of which the value of the test is estimated or judged. The criteria for evaluating a general intelligence examination, for example, may be school marks, ratings for aptitude in learning, or some other test believed to be valid, such as Stanford-Binet. A trade test may be validated against demonstrated ability to carry on the required operations as shown in actual performance.* A high correlation between a test and a criterion is evidence of validity provided the test and the criterion are both reliable. But before accepting criterion correlations, we must know the reliability of the test and if possible the reliability of the criterion.

When a criterion is not immediately available, indirect methods may be utilized for estimating the validity of a test.

* Stead, W. H., and Shartle, C. L., *Occupational Counseling Techniques* (1940), Chapters 5 and 8 especially.

We may, for example, compute the average correlation which each test in a battery shows with all of the other tests, and estimate the validity (i.e., the representativeness) of each test by the size of its correlations. Again, following essentially the same method, we may combine the scores on a number of tests designed to measure the same function (memory, say), and consider as most valid that test which correlates highest with the average of them all. Anastasi,* for example, found that of eight tests of immediate memory, the paired-associates test (geometric form paired against numbers) had the largest average correlation (i.e., .49), with the other tests of the battery. This test, then, is the most valid measure of the function tapped in common by all of the tests.

2. The Correction for Attenuation

The correlation between a test and its criterion will be reduced if either the test scores or the criterion scores or both are unreliable. In order to estimate the correlation between true scores in two variables, we need to make a correction which will take account of the unreliability in both sets of measures. Such a correction is given by the formula

$$r_{\infty\infty} = \frac{r_{12}}{\sqrt{r_{1I} \times r_{2II}}} \tag{85}$$

(correlation between true measures in Tests 1 and 2)

in which

$r_{\infty\infty}$ = correlation between true scores in Tests 1 and 2;
r_{12} = correlation between obtained scores in Tests 1 and 2;
r_{1I} = reliability coefficient of Test 1;
r_{2II} = reliability coefficient of Test 2.

Formula (85) is the well-known correction for attenuation formula. It provides a correction for the effects of those chance or accidental errors in the two tests which lower the reliability

* Anastasi, A., *A Group Factor in Immediate Memory*, Archives of Psychology, No. 120 (1930), p. 41.

coefficients of both tests and thus affect the correlation between them. To illustrate the application of formula (85), let the obtained correlation between two tests A and B be .60, the reliability coefficient of Test A be .80 (r_{1I}) and the reliability coefficient of Test B be .90 (r_{2II}). What is the correlation between Tests A and B freed of chance errors? Substituting the given values in formula (85), we have

$$r_{\infty\infty} = \frac{.60}{\sqrt{.80 \times .90}} = .71$$

as the estimated correlation between true scores in A and B. Our corrected coefficient of correlation represents the relationship which we should expect to obtain if our two sets of test scores were perfect measurements.

It is clear from formula (85) that correcting for chance errors will always raise the correlation between two tests — unless the reliability coefficients are both 1.00. Chance errors, therefore, always lower or attenuate an obtained correlation coefficient. The expression $\sqrt{r_{1I} \times r_{2II}}$ sets an upper limit to the correlation which we can obtain between two tests as they stand. In the example above, $\sqrt{.80 \times .90} = .85$; hence, Tests A and B cannot correlate higher than .85, as otherwise their corrected r would be greater than 1.00.

Let us assume the correlation between first year college grades and a general intelligence test to be .46; the reliability of the intelligence test to be .82; and the reliability of college grades to be .70. The maximum correlation which we could hope to obtain between these two measures is $\frac{.46}{\sqrt{.70 \times .82}}$ or .60. Knowing that the correlation between grades and general intelligence, corrected for errors of measurement, has a probable maximum value of .60 gives us a better notion of the "intrinsic" relationship between the two variables. At the same time, the investigator should remember that the $r_{\infty\infty}$ of .60 is a theoretical, not an obtained, value; that it gives an estimate of the relationship to be expected when the tests are more effective than they

actually were in the present instance. If many sources of error are present so that considerable correction is necessary, it would be better experimental technique to improve the tests and the experimental conditions than to correct the obtained r.

The investigator must be careful how he applies formula (85) to correlations which have been averaged, as in such cases the reliability coefficients may be lower than the correlations between the two tests. When this happens $r_{\infty\infty}$ is greater than 1.00. Such a result is logically and psychologically meaningless. If a corrected r is 1.00, or is only slightly greater than 1.00, however, it may be taken as indicating complete agreement between the two variables within the error of computation.

3. The Estimation of the True σ of a Test

Chance or variable errors have a marked effect upon the standard deviation of a test, as well as upon the r between tests. The relation of the σ calculated from obtained scores on a test to the σ of true scores on the same test is given by the formula

$$\sigma_\infty = \sigma_1\sqrt{r_{1I}} \tag{86}$$

(relation between true and obtained σ's for a set of test scores)

in which

σ_∞ = the σ of the true test scores;
σ_1 = the σ of the obtained test scores;
r_{1I} = the reliability coefficient of the test.

Suppose an educational achievement test of seventy-five items has been administered to a group of fifty children. The obtained standard deviation, σ_1, is 10, and the reliability coefficient of the test (r_{1I}) is .50. What is σ_∞, the σ of the true scores from which variable or accidental errors have been eliminated? Substituting $\sigma_1 = 10$, and $r_{1I} = .50$ in formula (86)

$$\sigma_\infty = 10\sqrt{.50}$$
$$= 7.1$$

and the "true σ" of the test is about 7 points.

It is clear from (86) that σ_∞ will *always* be smaller than σ_1,

except in the improbable case in which $r_{1I} = 1.00$. The effect of chance errors of measurement, then, is always to increase the spread (σ_1) of obtained test scores or of criterion scores.

4. Validation of a Test Battery*

A criterion of job efficiency, say, or of success in salesmanship may be forecast by a battery consisting of four, five, or more tests. The validity of such a battery is determined by the multiple correlation coefficient, R, between the battery and the criterion. The weights to be attached to scores on the sub-tests of the battery are given directly by the regression coefficients (p. 421).

If the regression weights are small fractions (as they often are) whole numbers may be substituted for them with little if any loss in accuracy. For example, suppose that the regression equation joining the criterion and the tests in a battery reads as follows:

$$C \text{ (criterion)} = 4.32X_1 + 3.12X_2 - .65X_3 + 8.35X_4 + K \text{ (constant)}$$

Dropping fractions and taking the nearest whole numbers, we have

$$C = 4X_1 + 3X_2 - 1X_3 + 8X_4 + K$$

Scores in Test 1 should be multiplied by 4, scores in Test 2 by 3, scores in Test 3 by -1, and scores in Test 4 by 8, in order to provide the best forecast of C, the criterion. The fact that Test 3 has a negative weight does not mean that this test has no value in forecasting C, but simply that the best estimate of C is obtained by giving scores in Test 3 a negative value.

III. Item Analysis

In Section II above, we considered the validity of final test scores. The validity of a test score also depends directly upon the care with which the *items* in the test have been chosen. While the subject of item analysis properly belongs in a book on

* See Chapter XIII.

test construction, the main features of the process may be outlined here. Item analysis may be divided into three main topics: (1) item selection, (2) item difficulty, and (3) item validity.

1. Item Selection

The initial choice of test items depends upon the judgment of competent persons as to the suitability of the material for the purposes of the test. Certain types of items, for instance, have proved to be generally useful in intelligence examinations. Problems in mental arithmetic, for example, vocabulary, analogies, and number series completion, are often encountered; also, items requiring generalization, interpretation and the ability to see relations. The validity of most standard tests of educational achievement depends upon the consensus of teachers and other competent judges as to the adequacy of the items included. Courses of study, requirements for different grades, curricula from different sections of the country are carefully culled over by the test makers to determine what material in history, English, geography, etc., should be included in an educational achievement battery designed, say, for the seventh grade. In its final form the educational achievement test represents items carefully selected from all available sources of information.

Items used in personal data sheets, interest inventories, attitude scales and the like, also represent a consensus of experts as to the most diagnostic items in the areas sampled.

2. Item Difficulty

The difficulty of an item is determined by the proportion of some standard group able to solve the item correctly. The scaling of separate test items has been described in Chapter VI, page 146. When normality of distribution can be assumed for the ability being measured, single items or groups of items (scores) may be scaled, i.e., given difficulty values along a scale in terms of σ. It has been customary to select items for a test

which vary in difficulty from easy to hard. The average person in the standardization group will then pass about one-half (50%) of the items in the test. It can be shown, however, that the sharpest discrimination as between good and poor subjects is provided by items which are passed by 50% of the members of a group. A test made up of items all of which are passed by approximately 50% (but by different persons, of course) would theoretically be the most discriminating test. But it would be difficult to construct such an examination and it is probable that a test made up of items covering a wider range of difficulty is psychologically a better measuring device. In standardizing a test care must be taken that few, if any, subjects achieve perfect or zero scores, as in neither case is the person measured by the test.

3. Item Validity

An often-used method of validating a test item is to determine whether the item discriminates between subjects differing sharply in the function being measured. This "criterion of internal consistency" admits into the final test or questionnaire only those items which have been found to separate high-scoring and low-scoring members of the group. In an internally consistent test, items "hang together" in the sense that they work in the same direction and measure the same common trait.* In one study,† eighty-six items were selected out of 222 on the basis of their ability to discriminate among the lower, middle, and upper thirds of the group. These eighty-six "good" items did a better job (higher reliability and validity) than a test nearly three times longer.

The validity of a single test item may also be determined by finding its correlation with total scores in the test of which it is a part, or by finding its correlation with scores in some inde-

* Ferguson, G. A., "The Factorial Interpretation of Test Difficulty," *Psychometrika*, 6 (1941), 323–329.

† Anderson, J. E., "The Effect of Item Analysis upon the Discriminative Power of an Examination," *Journal of Applied Psychology*, 19 (1935), 237–244.

pendent criterion. The bi-serial method (p. 347) is the standard procedure for determining item validity through correlation. Application of bi-serial r to each item in a test requires considerable computation, however. For this reason various short-cut methods for selecting good items by formula and by graphical methods have been devised. References given below should be consulted.*

PROBLEMS

1. The reliability coefficient of a test is .60.
   (*a*) How much must this test be lengthened in order to raise the self-correlation to .90?
   (*b*) What effect will doubling the test's length have upon its reliability coefficient? tripling the test's length?
2. A test of fifty items has a reliability coefficient of .78. What is the reliability coefficient
   (*a*) of a test having 100 items comparable to the items in the given test?
   (*b*) of a test having 125 comparable items?
3. A given test has a reliability coefficient of .80 and a σ of 20.
   (*a*) What is the maximum correlation which this test is capable of yielding as it stands (see p. 391)?
   (*b*) What is the standard error of a score obtained on this test?
   (*c*) What is the estimated reliability coefficient of this test in a group in which the σ is 15?
4. A test of 100 items is given to a group of 225 subjects with the following results: $M = 62.50$; $\sigma = 9.62$.
   (*a*) What is the reliability coefficient of the test by formula (79)?
   (*b*) What is the estimated true σ of this test?
   (*c*) What is the standard error of a score on this test?

* Long, John A., and Sandiford, Peter, *The Validation of Test Items*, Bulletin 3, 1935, University of Toronto, Department of Educational Research.
Flanagan, J. C., *General Considerations in the Selection of Test Items*, *Journal of Educational Psychology*, 30 (1939), 674–680.
Guilford, J. P., *The Phi-coefficient and Chi-square as Indices of Item Validity*, Psychometrika, 6 (1941), 11–19.
Richardson, M. W., and Adkins, D. C., *A Rapid Method of Selecting Test Items*, *Journal of Educational Psychology*, 29 (1928), 547–552.
Hawkes, H. E., Lindquist, E. R., and Mann, C. R., *Achievement Examinations*, 1936, Chaps. 2 and 3 especially.

5. Show (*a*) that when the reliability coefficient is zero, the standard error of an obtained score equals the standard deviation of the test; and (*b*) that when the reliability coefficient is 1.00, the standard error of an obtained score equals zero.
6. A mathematics test has a reliability coefficient of .82, and a mechanical ability test has a reliability coefficient of .76. The r between the two tests is .52.
   (*a*) What would the correlation be if *both* tests were perfect measures?
   (*b*) What is the maximum correlation possible with the mathematics test as it stands?
   (*c*) What is the maximum correlation possible with the mechanical ability test as it stands?
7. An intelligence examination shows a correlation of .50 with first-year scholarship. The reliability coefficient of the test is .85, and of school grades (i.e., the criterion) is .65. What is the highest validity coefficient which we can hope to get with this test (i.e., corrected correlation between test and grades)?
8. A test of seventy-five items has a σ_t of 12.35. The $\Sigma pq = 16.46$. What is the reliability coefficient by formula (78)?

ANSWERS

1. (*a*) six times
   (*b*) $r_{1I} = .75$ (doubling length); $r_{1I} = .82$ (tripling length)
2. (*a*) .88
   (*b*) .90
3. (*a*) .89
   (*b*) 8.9
   (*c*) .64
4. (*a*) .75
   (*b*) 8.34
   (*c*) 4.81
6. (*a*) .66
   (*b*) .91
   (*c*) .87
7. .68
8. .90

CHAPTER XIII

PARTIAL AND MULTIPLE CORRELATION

I. The Meaning of Partial and Multiple Correlation

Partial and multiple correlation represent an important extension of the theory and technique of simple or two-variable correlation to problems which involve three or more variables. In computing the correlation between two sets of scores, it is often desirable to allow for the influence of factors which through their common relationship to the variables being correlated obscure results or make them difficult to interpret. To illustrate, suppose that the correlation between intelligence test scores and chronological age in a large group of children, seven to fourteen years old, is .50; that the correlation between school achievement and age in the same group is .40; and that the correlation between intelligence and school achievement is .70. Since intelligence test scores and school achievement both increase with age (the correlations are .50 and .40) the correlation between these two measures will be raised when age is allowed to vary. The correlation coefficient of .70, therefore, is not only a measure of the role of intelligence in school achievement, but is a measure of the influence of intelligence *plus* the indirect effects of differences in age or maturity upon school achievement.

To discover the relationship between intelligence and school achievement, uninfluenced by maturity, we must rule out or control the factor of age. This could be accomplished *experimentally* by selecting children all of whom are of the same age. But this procedure offers many difficulties, the principal one being that it is well-nigh impossible to find a large sample of children of exactly the same age. It becomes necessary, then, to determine what age range is permissible; and the more

closely we limit our group with respect to age, the smaller the number left. In fact, the experimental control of a variable by the method of selection may so limit the size of the group that correlations are of doubtful value.

Because of the difficulties which arise in attempting to control a variable (or variables) experimentally, the method of partial correlation is often employed. By this method the relationship between two variables can be determined when one or more related variables are held constant. Thus, the partial correlation between general intelligence and school achievement, i.e., the correlation with age "partialled out," gives us the correlation between these two variables uninfluenced by the factor of age differences. Such a partial coefficient represents the *net* correlation between general intelligence and school achievement for children of the *same age;* or the net correlation between intelligence and school achievement when age is a constant factor. Expressed in still another way, our partial coefficient tells us what relationship exists between general intelligence test scores and school achievement when differences in maturity no longer affect *either* variable.

A second illustration of partial correlation may be helpful. A teacher finds in her class a correlation of .60 between test scores in history and arithmetic. In looking for an explanation of this correlation (since there is apparently little reason to *expect* a high relationship between these two abilities), she finds that achievement in arithmetic seems to depend in part upon ability to read and understand the problems. Obviously, ability to read well is also an important factor in determining achievement in history. Suppose that our teacher now calculates the correlations of the history and arithmetic tests with a *third* test of reading comprehension. Knowing these r's, she may determine (by methods given on p. 414) the net or partial correlation between history and arithmetic when differences in reading comprehension have been allowed for. If this partial coefficient is .30, say — considerably smaller than the "whole" coefficient (of .60) between history and arithmetic — the hypothesis that

the apparent relationship was due in part to the common dependence of both tests upon reading is verified. When a factor (or factors) is "partialled out" from a given correlation the effect is to eliminate the differences among individuals introduced by the variable thus controlled. The method of eliminating factor variability through partial correlation may be employed whenever the correlation can be computed between the factor or factors to be controlled and the two variables the net correlation of which we are seeking. Since *all* of the data are utilized, partial correlation has a decided advantage over experimental control in many problems.

In addition to its value as a means of controlling conditions by eliminating the effects of "disturbing" or other variables, partial correlation is useful in other ways. It enables us, for example, to build up a regression equation involving three or more variables from which a "criterion" score may be predicted when we know the scores made by a subject on several correlated tests. The accuracy of the regression equation in estimating criterion scores — its reliability as a "prediction" instrument — can be determined by the *multiple coefficient of correlation.* A multiple correlation coefficient gives the correlation between a single test or criterion on the one hand and a *team* of tests on the other. The meaning of the multiple coefficient of correlation will be better understood when the student has worked through an actual problem such as that given in Table 59.

II. An Illustrative Correlation Problem Involving Three Variables

Perhaps the most straightforward approach to an understanding of the meaning of partial and multiple correlation, and of the techniques of calculation involved, is through the solution of a problem. The present section, therefore, will show the application of partial and multiple correlation to a three-variable problem. Following this, the general formulas and further applications of the method will be considered.

TABLE 59

A CORRELATION PROBLEM INVOLVING THREE VARIABLES

(To illustrate partial and multiple correlation)

Step 1. Primary Data ($N = 450$)

(1) Honor Points	(2) General Intelligence	(3) Average Hours of Study per Week
$M_1 = 18.5$	$M_2 = 100.6$	$M_3 = 24$
$\sigma_1 = 11.2$	$\sigma_2 = 15.8$	$\sigma_3 = 6$
$r_{12} = .60$	$r_{13} = .32$	$r_{23} = -.35$

Step 2. Calculation of Partial Coefficients of Correlation

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r^2_{13}}\sqrt{1 - r^2_{23}}} = \frac{.60 - .32(-.35)}{.9474 \times .9367} = .80 \quad (87)$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r^2_{12}}\sqrt{1 - r^2_{23}}} = \frac{.32 - .60(-.35)}{.8000 \times .9367} = .71 \quad (87)$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r^2_{12}}\sqrt{1 - r^2_{13}}} = \frac{(-.35) - .60 \times .32}{.8000 \times .9474} = -.72 \quad (87)$$

Step 3. The Regression Equations and Partial Regression Coefficients

$$\bar{x}_1 = b_{12.3}x_2 + b_{13.2}x_3 \quad \text{(Deviation Form)} \quad (89)$$

or

$$\bar{X}_1 = b_{12.3}X_2 + b_{13.2}X_3 + K \quad \text{(Score Form)} \quad (90)$$

in which

$$b_{12.3} = r_{12.3}\frac{\sigma_{1.23}}{\sigma_{2.13}} \text{ and } b_{13.2} = r_{13.2}\frac{\sigma_{1.23}}{\sigma_{3.12}} \quad (93)$$

Step 4. Calculation of the Partial σ's

(1) $\sigma_{1.23} = \sigma_1\sqrt{1 - r^2_{12}}\sqrt{1 - r^2_{13.2}} = 11.2 \times .8000 \times .7042 = 6.3$ (88)

(2) $\sigma_{2.13} = \sigma_2\sqrt{1 - r^2_{23}}\sqrt{1 - r^2_{12.3}} = 15.8 \times .9367 \times .6000 = 8.9$ (88)

(3) $\sigma_{3.12} = \sigma_3\sqrt{1 - r^2_{23}}\sqrt{1 - r^2_{13.2}} = 6 \times .9367 \times .7042 = 4.0$ (88)

Step 5. Calculation of the Partial Regression Coefficients, and Partial Regression Equation

Substituting for $r_{12.3}$, $r_{13.2}$, $\sigma_{1.23}$, $\sigma_{2.13}$, $\sigma_{3.12}$, we have

$$b_{12.3} = .80 \times \frac{6.3}{8.9} = .57; \; b_{13.2} = .71 \times \frac{6.3}{4.0} = 1.12$$

Hence the regression equation becomes:

$$\bar{x}_1 = .57x_2 + 1.12x_3 \quad \text{(Deviation Form)}$$

or

$$\bar{X}_1 = .57X_2 + 1.12X_3 - 66 \quad \text{(Score Form)}$$

Step 6. Calculation of the Standard Error of Estimate

$$\sigma_{(\text{est. } X_1)} = \sigma_{1.23} = 6.3 \quad (96)$$

Step 7. Calculation of the Coefficient of Multiple Correlation

$$R_{1(23)} = \sqrt{1 - \frac{\sigma^2_{1.23}}{\sigma^2_1}} \quad (98)$$

$$= .83$$

The problem in Table 59 is taken from a study* of the factors which influence "academic success." In that part of the study from which the present data are drawn, the problem was to discover how accurately one can predict the academic success of freshmen from a knowledge of their general intelligence and of their study habits. Academic success was defined specifically as the number of credit or "honor" points obtained by a student at the end of his first semester in college. The number of honor points earned depended upon the number of *A*, *B*, and *C* grades made by the student in his freshman courses. A grade of *A* carried three honor points; a grade of *B* two honor points; a grade of *C* one honor point; and a grade of *D*, which was a passing mark, carried no honor point credit. The maximum number of points which a freshman taking the regulation number of courses in one semester could obtain was forty-eight.

General intelligence was measured by a combination of the Miller Mental Ability Test, and the Dartmouth Completion of Definitions Test. The first test contains 120 items and the second 40, so that the maximum score was 160. The scores of the 450 students in this sample ranged from 50 to 150, the distribution being fairly normal. As a measure of interest and application it was decided to take the average number of hours per week spent in study. Information with regard to study habits was obtained by means of a questionnaire given at the beginning and again at the middle of the first semester. Among other items in the questionnaire upon which information was requested were the number of hours spent per week at meals, in sleeping, etc. These and other questions were included in order that the student might think that he was being checked upon the distribution of his total time and not upon his study habits alone. The correlation between the student's estimates of the number of hours spent in study (given on the first and second questionnaires) was .86, indicating a satisfactory degree of reliability.

As stated above, the main object of this study was to find how

* May, M. A., "Predicting Academic Success," *Journal of Educational Psychology*, 14 (1923), 429–440.

accurately the number of honor points which a student earns can be predicted from a knowledge of his study habits and his general intelligence. Other factors, of course, such as health, personality, previous preparation, and the like, are undoubtedly of importance in determining the number of honor points received. The two factors selected were chosen because they are important and are also objective and measurable. As the first step in solving our problem, we shall calculate the partial coefficient which shows to what extent honor points are related to general intelligence when the variable factor of study hours per week is held constant. Next the partial coefficient will be calculated which shows to what extent honor points are related to study hours when the variable effect of general intelligence is rendered constant. Apart from the employment of these partial coefficients in the regression equation from which we predict honor points, the information which they yield will prove in itself to be of considerable interest. The solution of the problem is outlined in the following series of steps; the necessary data and calculations will be found in Table 59.

Step 1

The mean and σ of each series of measures and the intercorrelations are first calculated. These intercorrelations are product-moment r's computed as shown in Chapter IX. The correlation between (1) honor points and (2) general intelligence, written r_{12}, is .60; the correlation between (1) honor points and (3) the number of hours per week spent on the average in study, written r_{13}, is .32; and the correlation between (2) general intelligence and (3) hours of study per week, written r_{23}, is $-.35$. The low correlation between honor points and study hours is of decided interest; but the most surprising correlation is the $-.35$ between study hours and general intelligence. Evidently the brighter the student, the less he studies.

Step 2

Having found the intercorrelations of our three variables, we may then calculate the net correlation between (1) honor points

and (2) general intelligence with the influence of (3) study hours partialled out or held constant. This net or partial coefficient of correlation, written $r_{12.3}$, is found from the following formula:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r^2_{13}}\sqrt{1 - r^2_{23}}} \qquad \text{(87), p. 415}$$

Substitution of the values for r_{12}, r_{13}, and r_{23} in the formula gives a partial coefficient, $r_{12.3}$, of .80. This means that if *all* of our 450 students had studied exactly the same number of hours per week, the coefficient of correlation between honor points earned and general intelligence test scores would have been .80 instead of .60. In other words, if each student spends the *same number of hours in study*, there is a closer correspondence between general intelligence test scores and honor points earned than there is when the number of study hours varies.

The partial coefficient of correlation between (1) honor points and (3) hours spent in study per week with (2) general intelligence partialled out, or its influence held constant, is found from the formula

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r^2_{12}}\sqrt{1 - r^2_{23}}} \qquad (87)$$

Substitution of the values for r_{13}, r_{12}, and r_{23} gives a partial coefficient, $r_{13.2}$, of .71, as against an obtained coefficient (r_{13}) of .32. This result means that if our group possessed the same general intelligence* there would be a much closer correspondence between the number of honor points received and the number of hours spent in study than there is when the members of the group possess varying degrees of intelligence. This is certainly the result to be expected.

The last partial coefficient of correlation $r_{23.1}$ equals − .72. This coefficient gives the net correlation between (2) general intelligence and (3) study hours when the influence of (1) honor points is held constant. It is found from the formula

* By "same general intelligence" is meant the same *score* on the given general intelligence tests.

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1 - r^2_{12}}\sqrt{1 - r^2_{13}}} \tag{87}$$

Like the two partial r's above, we may interpret $r_{23.1}$ to mean that the correlation between general intelligence and hours spent in study in a group in which every student earns the same number of honor points would be much higher (in the *inverse* direction) than the "raw" correlation between the same two factors in an unselected group. By an unselected group is meant here a group in which the number of honor points received by different students varies. It seems evident that the brighter student not only studies less than the average and dull (since $r_{23} = -.35$) but that the brighter the student, the less he *needs* to study in order to reach a given standard of academic success — earn a given number of honor points.

Step 3

Knowing the partial coefficients of correlation, we may write the multiple regression equation from which the *most probable* number of honor points a student will receive may be estimated when we know his score in the general intelligence test and the number of hours he studies per week. The regression equation for three variables (in *deviation form*) is as follows:

$$\bar{x}_1 = b_{12.3}x_2 + b_{13.2}x_3 \qquad \text{(89), p. 419}$$

In this equation x_1 stands for honor points and is the *dependent* variable or criterion; x_2 and x_3 stand for general intelligence and study hours, respectively, and are the *independent* variables. Note the resemblance of this equation to the simple regression equation for two variables $\bar{y} = b_{12} \times x$ (p. 312). If x_1 is put for $\bar{y}$, and x_2 for x in the two-variable equation, we have $x_1 = b_{12} \times x_2$.

When written in *score form*, the multiple regression equation for three variables becomes

$$(X_1 - M_1) = b_{12.3}(X_2 - M_2) + b_{13.2}(X_3 - M_3)$$

or transposing and collecting terms,

$$\bar{X}_1 = b_{12.3}X_2 + b_{13.2}X_3 + K \text{ (a constant)} \qquad \text{(90), p. 419}$$

It is clear that before we can use this equation we must find the value of the *partial regression coefficients* $b_{12.3}$ and $b_{13.2}$. These may be found from the formulas

$$b_{12.3} = r_{12.3}\frac{\sigma_{1.23}}{\sigma_{2.13}} \text{ and } b_{13.2} = r_{13.2}\frac{\sigma_{1.23}}{\sigma_{3.12}} \qquad (93), \text{ p. } 420$$

and, as we already have the values of $r_{12.3}$ and $r_{13.2}$, it is only necessary that we find $\sigma_{1.23}$, $\sigma_{2.13}$, and $\sigma_{3.12}$ (the partial σ's) in order to replace the partial regression coefficients in the equation by numerical values.

Note that the partial coefficient of correlation $r_{23.1}$, although of interest as giving us the relation between general intelligence and hours spent in study for a constant number of honor points earned, is not actually needed in the regression equation $\bar{x}_1 = b_{12.3}x_2 + b_{13.2}x_3$. In order to evaluate the constants $b_{12.3}$ and $b_{13.2}$ in our regression equation, we need *only* $r_{12.3}$ and $r_{13.2}$. In fact, in *any* problem involving three variables, only *two* partial coefficients of correlation need be computed, if we are interested primarily in the prediction of X_1 scores from known values of X_2 and X_3.

Step 4

The partial σ's may be found from the formulas

$$\begin{aligned} \sigma_{1.23} &= \sigma_1\sqrt{1 - r^2_{12}}\,\sqrt{1 - r^2_{13.2}} \\ \sigma_{2.13} = \sigma_{2.31} &= \sigma_2\sqrt{1 - r^2_{23}}\,\sqrt{1 - r^2_{12.3}} \qquad (88), \text{ p. } 417 \\ \sigma_{3.12} = \sigma_{3.21} &= \sigma_3\sqrt{1 - r^2_{23}}\,\sqrt{1 - r^2_{13.2}} \end{aligned}$$

Substituting the known values of the raw and partial r's in these formulas we find that $\sigma_{1.23} = 6.3$; $\sigma_{2.13} = 8.9$; and $\sigma_{3.12} = 4.0$. (For the calculations see Table 59).

Step 5

From the partial σ's and the partial r's the numerical values of the partial regression coefficients $b_{12.3}$ and $b_{13.2}$ are found to be .57 and 1.12, respectively. We may now write the multiple regression equation in deviation form as

$$\bar{x}_1 = .57x_2 + 1.12x_3$$

In order to write this multiple regression equation in score form we replace x_1 by $(X_1 - 18.5)$; x_2 by $(X_2 - 100.6)$; and x_3 by $(X_3 - 24)$. The equation then becomes

$$\overline{X}_1 = .57X_2 + 1.12X_3 - 66$$

Given a student's general intelligence test score (X_2) and the number of hours per week he spends in study (X_3), we can estimate from this equation the "most probable" number of honor points he will receive during his first semester in college. Suppose that student J. N. has a general intelligence test score of 120 and that he studies on the average twenty hours per week: how many honor points will he most probably receive during the first semester? Substituting $X_2 = 120$ and $X_3 = 20$ in the regression equation, we find that

$$\overline{X}_1 = (.57 \times 120) + (1.12 \times 20) - 66 = 25$$

The most probable number of honor points which student J. N. will receive, therefore, using the given measures as the basis of our forecast, is twenty-five.

Step 6

This forecast, like every other "most probable" number of honor points predicted from the regression equation, has an "error of estimate." The standard error of estimate of *any* X_1 predicted from the regression equation, $\overline{X}_1 = b_{12.3}X_2 + b_{13.2}X_3 + K$ is written $\sigma_{(\text{est}.X_1)}$, and equals $\sigma_{1.23}$ directly (p. 418).

The standard error of estimate in the present problem is 6.3, and in the illustration given above, the twenty-five honor points estimated for J. N. have a $SE_{(\text{est}.X_1)}$ of about six points. This means that the chances are about two in three that our forecast of twenty-five honor points will not miss the actual number of honor points received by J. N. by more than $\pm$ 6. In general we may say that two-thirds of all predicted honor point values will lie within $\pm$ 6 points of their actual values.

Step 7

The final step in the solution of our three-variable correlation problem is the computation of the coefficient of multiple cor-

relation. "Multiple r," generally written R, is defined (see p. 425) as the coefficient of correlation between scores *actually made* on the criterion test and scores on the same test *predicted* from the regression equation. In the present problem, R gives the correlation between *earned* honor points (X_1) and honor points *estimated* by means of the two variables, general intelligence (X_2) and hours of study (X_3), when these two are *combined* into a team by means of the regression equation. The formula for R when we are dealing with three variables is

$$R_{1(23)} = \sqrt{1 - \frac{\sigma^2_{1.23}}{\sigma^2_1}} \qquad \text{(97), p. 424}$$

In the present problem $R_{1(23)} = .83$. This means that if the most probable number of honor points which each student in our group of 450 will receive is predicted from the regression equation given on page 413, the correlation between these 450 *predicted scores* and the 450 *scores actually received* will be .83. Multiple R tells us to what extent X_1 is determined by the combined action of X_2 and X_3; or, in the present instance, to what extent honor points are related to general intelligence and number of study hours per week taken together.

III. General Formulas for Use in Partial and Multiple Correlation

1. Partial r's of Any Order

(1) Formulas for Partial r's

We found in Table 59 that one is able by the method of partial correlation to find the net relationship between two variables when the influence of a third is ruled out or held constant. By an extension of the partial correlation method, we may obtain the net correlation between X_1 and X_2 when *two* or *more* variables have been held constant. The partial coefficient of correlation $r_{12.34}$, for example, means by analogy to $r_{12.3}$ that the correlation between X_1 and X_2 has been freed of the influence of *both* X_3 and X_4; and the partial coefficient of correlation $r_{12.34\ldots n}$

means that the correlation between X_1 and X_2 has been freed of the influence of a large number of disturbing factors.

In every partial coefficient of correlation, e.g., $r_{12.34}$, the *primary* subscripts to the *left* of the point (1 and 2) define the two variables whose net correlation we are seeking. The *secondary* subscripts to the *right* of the point (3 and 4) denote the variables ruled out or held constant. The *order* in which the secondary subscripts are written is immaterial, i.e., $r_{12.34} = r_{12.43}$. The order of the primary subscripts is of importance, however, as it tells us which variable is taken to be dependent and which independent. Thus r_{12} means that X_1 is dependent — is to be predicted from X_2; while r_{21} means that X_2 is dependent — is to be predicted from X_1. The numerical values r_{12} and r_{21} are, of course, the same. The order of a partial r is determined by the *number* of its secondary subscripts. Thus r_{12}, an "entire" or "total" r, is a coefficient of zero order; $r_{12.3}$ is a partial r of the *first* order; $r_{12.345}$ is a coefficient of the *third* order.

The general formula for a partial r is

$$r_{12.34\,.\,.\,n} = \frac{r_{12.34\,.\,.\,.\,(n-1)} - r_{1n.34\,.\,.\,.\,(n-1)}\,r_{2n.34\,.\,.\,.\,(n-1)}}{\sqrt{1 - r^2_{1n.34\,.\,.\,.\,(n-1)}}\,\sqrt{1 - r^2_{2n.34\,.\,.\,.\,(n-1)}}} \tag{87}$$

(partial correlation coefficient in terms of the coefficients of lower order — n *variables)*

From this formula partial r's of any given order may be found. In a five-variable problem, for example, $(n - 1) = 4$, and $n = 5$, so that $r_{12.345}$ is written

$$r_{12.345} = \frac{r_{12.34} - r_{15.34}\,r_{25.34}}{\sqrt{1 - r^2_{15.34}}\,\sqrt{1 - r^2_{25.34}}}$$

that is, in terms of the partial r's of the *second* order. These second order partial r's must then be computed by formula (87) from r's of the *first* order before the third order r, $r_{12.345}$, can be evaluated. In calculating partial r's Table 60 may be used to read $\sqrt{1 - r^2}$ values.

There are several methods akin to partial correlation which are useful in certain special problems. Two of these, *part cor-*

TABLE 60

A TABLE TO INFER THE VALUE OF $\sqrt{1 - r^2}$ FROM A GIVEN VALUE OF r

r	$\sqrt{1 - r^2}$	r	$\sqrt{1 - r^2}$	r	$\sqrt{1 - r^2}$
.0000	1.0000	.3400	.9404	.6800	.7332
.01	.9999	.35	.9367	.69	.7238
.02	.9998	.36	.9330	.70	.7141
.03	.9995	.37	.9290	.71	.7042
.04	.9992	.38	.9250	.72	.6940
.05	.9987	.39	.9208	.73	.6834
.06	.9982	.40	.9165	.74	.6726
.07	.9975	.41	.9121	.75	.6614
.08	.9968	.42	.9075	.76	.6499
.09	.9959	.43	.9028	.77	.6380
.10	.9950	.44	.8980	.78	.6258
.11	.9939	.45	.8930	.79	.6131
.12	.9928	.46	.8879	.80	.6000
.13	.9915	.47	.8827	.81	.5864
.14	.9902	.48	.8773	.82	.5724
.15	.9887	.49	.8717	.83	.5578
.16	.9871	.50	.8660	.84	.5426
.17	.9854	.51	.8617	.85	.5268
.18	.9837	.52	.8542	.86	.5103
.19	.9818	.53	.8480	.87	.4931
.20	.9798	.54	.8417	.88	.4750
.21	.9777	.55	.8352	.89	.4560
.22	.9755	.56	.8285	.90	.4359
.23	.9732	.57	.8216	.91	.4146
.24	.9708	.58	.8146	.92	.3919
.25	.9682	.59	.8074	.93	.3676
.26	.9656	.60	.8000	.94	.3412
.27	.9629	.61	.7924	.95	.3122
.28	.9600	.62	.7846	.96	.2800
.29	.9570	.63	.7766	.97	.2431
.30	.9539	.64	.7684	.98	.1990
.31	.9507	.65	.7599	.99	.1411
.32	.9474	.66	.7513	1.00	.0000
.33	.9440	.67	.7424		

relation and *semi-partial correlation,* may be mentioned briefly. These procedures differ from partial correlation in that they give the net effect secured by ruling out the influence of one or more variables from only *one* of the two correlated measures, instead of from *both.* For example, one may wish to know the relation (semi-partial) between reaction time and speed of reading when differences in size of vocabulary are held constant with respect to reading only. Part correlation and semi-partial correlation

have not been widely used in mental measurement. For a discussion of formulas and for illustrations see references below.*

(2) Significance of a Partial r

The significance of a partial r (like that of a zero-order r) may be tested against the null hypothesis. We may use either Table 49, page 299 or Table 61, column headed "2 variables." The degrees of freedom for a partial r are $N - m$ where N = number of cases, and m = number of variables entering into the partial r. Thus if $r_{12.345} = .40$ and $N = 75$, $m = 5$ and $N - m = 75 - 5$ or 70.

In Table 59, $r_{12.3} = .80$, $N = 450$, $m = 3$, and $N - m = 447$. From Table 61, column 2, the r entries by interpolation for $N = 447$ are .093 and .121 at the .05 and .01 levels. The probability that the obtained $r_{12.3}$ of .80 arose from fluctuations of sampling is much less than .01; and this is true, also, of $r_{13.2}$ of .71 and $r_{23.1}$ of $-.72$. All three partial r's, in fact, are highly significant.

2. Partial σ's of Any Order

General Formulas

Just as the correlation between two sets of scores can be determined when the influence of 1, 2, 3 . . . n factors is held constant, so the variability (σ) of a set of scores can be computed when the influence of 1, 2, 3 . . . n variables is ruled out. As an illustration, consider $\sigma_{1.23}$ of Table 59. This partial σ gives the variability of X_1 (honor points) *freed* of the influence upon variability exerted by the two factors X_2 (general intelligence) and X_3 (study hours per week). The general formula for partial σ's of any order is

$$\sigma_{1.234\ldots n} = \sigma_1\sqrt{1 - r^2_{12}}\sqrt{1 - r^2_{13.2}}\sqrt{1 - r^2_{14.23}} \ldots \sqrt{1 - r^2_{1n.23\ldots(n-1)}} \qquad (88)$$

(*partial* σ *for* n *variables*)

* Ezekiel, M., *Methods of Correlation Analysis* (2nd ed., 1941), p. 213. Dunlap, J. W., and Cureton, E. E., "On the Analysis of Causation," *Journal of Educational Psychology*, 21 (1930), 657–680.

This formula may be used to compute the net σ's in correlation problems which involve any number of variables. In a five-variable problem, for example, $\sigma_{1.2345}$ is written

$$\sigma_{1.2345} = \sigma_1\sqrt{1 - r^2_{12}}\,\sqrt{1 - r^2_{13.2}}\,\sqrt{1 - r^2_{14.23}}\,\sqrt{1 - r^2_{15.234}}$$

This partial σ is of the *fourth* order since it has four secondary subscripts, and the order of a partial σ, like the order of a partial r, is determined by the number of its secondary subscripts.

By a simple rearrangement of the secondary subscripts, any higher order σ may be written in more than one way. A partial σ of the second order may be written in two ways: for example, $\sigma_{1.23}$ which is given on page 412 as

$$\sigma_{1.23} = \sigma_1\sqrt{1 - r^2_{12}}\,\sqrt{1 - r^2_{13.2}}$$

may also be written

$$\sigma_{1.32} = \sigma_1\sqrt{1 - r^2_{13}}\,\sqrt{1 - r^2_{12.3}}$$

In like manner $\sigma_{2.13}$ may be written

$$(1) \qquad \sigma_{2.13} = \sigma_2\sqrt{1 - r^2_{12}}\,\sqrt{1 - r^2_{23.1}}$$

or

$$(2) \qquad \sigma_{2.31} = \sigma_2\sqrt{1 - r^2_{23}}\,\sqrt{1 - r^2_{12.3}}$$

and $\sigma_{3.12}$ may be written

$$(1) \qquad \sigma_{3.12} = \sigma_3\sqrt{1 - r^2_{13}}\,\sqrt{1 - r^2_{23.1}}$$

or

$$(2) \qquad \sigma_{3.21} = \sigma_3\sqrt{1 - r^2_{23}}\,\sqrt{1 - r^2_{13.2}}$$

These alternate forms of a partial σ are useful as a check upon arithmetic calculations; also they make unnecessary the calculation of unused partial r's. Use of the *second* forms of $\sigma_{2.13}$ and $\sigma_{3.12}$ instead of the *first* (see Table 59), for example, makes it unnecessary to compute $r_{23.1}$ so far as the partial σ's in the regression equation are concerned. Furthermore, if $r_{23.1}$ is not wanted for other purposes, it need not be calculated at all (see p. 412). Two partial r's are all that are required in order to write the regression equation of a three-variable problem.

3. Multiple Regression Equations and Partial Regression Coefficients

(1) The Multiple Regression Equation for Any Number of Variables

The regression equation which expresses the relationship between a single dependent or criterion variable, X_1, and any number of independent variables, $X_2, X_3, X_4 \ldots X_n$ may be written in *deviation form* as follows:

$$\bar{x}_1 = b_{12.34 \ldots n}x_2 + b_{13.24 \ldots n}x_3 + \cdots + b_{1n.23 \ldots (n-1)}x_n \quad (89)$$

(*regression equation, deviation form, for* n *variables*)

and in *score form*

$$\bar{X}_1 = b_{12.34 \ldots n}X_2 + b_{13.24 \ldots n}X_3 + \cdots + b_{1n.23 \ldots (n-1)}X_n + K \quad (90)$$

(*regression equation, score form, for* n *variables*)

The partial regression coefficients $b_{12.34 \ldots n}$, $b_{13.24 \ldots n}$, etc., give the *weights* to be attached to the scores of each independent variable when X_1 is to be estimated from all of these in combination. Furthermore, the regression coefficients give the weight which each variable exerts in determining X_1 when the influence of the other variables is excluded. Hence, we can tell from the regression equation just what role each of the several test variables plays in determining the score on Test 1, the test taken as the criterion.

(2) The Multiple Regression Equation for Three Variables (Special Form)

When a problem involves only three variables, the regression equation, as we have seen, is written

$$\bar{x}_1 = b_{12.3}x_2 + b_{13.2}x_3 \quad \text{(deviation form)}$$

If the partial r's and partial σ's are of no special interest, it is possible to express the equation above in a somewhat more convenient form for calculation, as follows:

$$\bar{x}_1 = \frac{\sigma_1(r_{12} - r_{13}r_{23})}{\sigma_2(1 - r^2_{23})} x_2 + \frac{\sigma_1(r_{13} - r_{12}r_{23})}{\sigma_3(1 - r^2_{23})} x_3 \tag{91}$$

(regression equation for three variables, special form)

or in score form

$$\bar{X}_1 = \frac{\sigma_1(r_{12} - r_{13}r_{23})}{\sigma_2(1 - r^2_{23})} X_2 + \frac{\sigma_1(r_{13} - r_{12}r_{23})}{\sigma_3(1 - r^2_{23})} X_3 + K \tag{92}$$

(regression equation for three variables, special form)

As this equation involves *only* zero order r's and zero order σ's, X_1 may be estimated from it without the computation of any partial r's or partial σ's. We may illustrate using the data given in Table 59, page 407. Substituting for $\sigma_1 = 11.2$, $\sigma_2 = 15.8$, $\sigma_3 = 6$, $r_{12} = .60$, $r_{13} = .32$, and $r_{23} = -.35$, we have

$$x_1 = \frac{11\ 2(.60 + .32 \times .35)}{15.8(1 - .35^2)} x_2 + \frac{11.2(.32 + .60 \times .35)}{6(1 - .35^2)} x_3$$

$$\bar{x}_1 = .57x_2 + 1.12x_3$$

which checks the regression equation as calculated in Table 59.

(3) Partial Regression Coefficients (b's)

Partial regression coefficients may be computed from the formula

$$b_{12.34 \ldots n} = r_{12.34 \ldots n} \frac{\sigma_{1.234 \ldots n}}{\sigma_{2.134 \ldots n}} \tag{93}$$

(partial regression coefficients in terms of partial coefficients of correlation and standard errors of estimate — n variables)

When the problem involves three variables, the regression coefficients, $b_{12.3}$ and $b_{13.2}$ are, like $r_{12.3}$ and $r_{13.2}$, of the *first* order. The first regression coefficient, $b_{12.3}$, equals $r_{12.3} \dfrac{\sigma_{1.23}}{\sigma_{2.13}}$ and the second regression coefficient, $b_{13.2}$, equals $r_{13.2} \dfrac{\sigma_{1.23}}{\sigma_{3.12}}$.

Partial regression coefficients which involve more than three variables may be calculated from formula (93). In a five-vari-

able problem, for example, the regression coefficients (of the *third* order) are

$$b_{12.345} = r_{12.345} \frac{\sigma_{1.2345}}{\sigma_{2.1345}}$$

$$b_{13.245} = r_{13.245} \frac{\sigma_{1.2345}}{\sigma_{3.1245}} \quad \text{etc.}$$

In order to find these partial regression coefficients we first compute the third order partial r's, and the fourth order partial σ's.

The b's are determined by the σ's of the tests and these in turn depend upon the units in terms of which the test is scored. The b-coefficients give the weights of *scores* in the independent variables, X_2, X_3, etc., but not the contribution of these variables without regard to the scoring system employed. The latter contribution is given by the "beta weights," described in (4) below.

(4) The Beta (β) Coefficients

When expressed in terms of standard or σ-scores, partial regression coefficients are usually called beta coefficients. The beta coefficients may be calculated directly from the b's as follows:

$$\beta_{12.34\ldots n} = b_{12.34\ldots n} \frac{\sigma_2}{\sigma_1} \tag{94}$$

(beta *coefficients calculated from partial regression coefficients*)

The multiple regression equation for n variables may also be written in standard scores as

$$\bar{z}_1 = \beta_{12.34\ldots n}z_2 + \beta_{13.24\ldots n}z_3 + \cdots + \beta_{1n.23\ldots(n-1)}z_n \tag{95}$$

(*multiple regression equation in terms of standard scores*)

Beta coefficients are often called "beta weights" to distinguish them from the "score weights" (b's) of the ordinary multiple regression equation. When all of our tests have been expressed in standard scores (all Means = .00 and all σ's = 1.00) differences in test units as well as differences in variability are allowed for. We are then able to determine from the correla-

tions alone the relative weight with which each independent variable "enters in" or contributes to the criterion, independently of the other factors.

To illustrate with the data in Table 59, we find that $\beta_{12.3} = .57 \times \frac{15.8}{11.2}$ or .81 and that $\beta_{13.2} = 1.12 \times \frac{6.0}{11.2}$ or .60. From (95) above we get

$$\bar{z}_1 = .81z_2 + .60z_3$$

This equation should be compared with the multiple regression equation $\bar{x}_1 = .57x_2 + 1.12x_3$ in Table 59 which gives the weights to be attached to the *scores* in X_2 and X_3. The weights of .57 and 1.12 tell us the amount by which scores in X_2 and X_3 must be multiplied in order to give the "best" prediction of X_1. But these weights do not give us the relative importance of general intelligence and study habits in determining the number of honor points a freshman will receive. This information is given by the beta weights. It is of interest to note that while the actual score weights are as 1:2 (.57 to 1.12), the independent contributions of general intelligence (z_2) and study habits (z_3) are in the ratio of .81 to .60 or as 4:3. When the variabilities (σ's) of our tests are all equal and scoring units are comparable, general intelligence has a proportionately greater influence than study habits in determining academic achievement. This is certainly the result to be expected.

4. The Standard Error of Estimate for Multiple Regression Equations

All X_1 scores estimated from a multiple regression equation have a standard error of estimate which measures the error made in taking scores given by the regression equation instead of *actual* scores (those earned on the criterion test). The standard error of estimate is given directly by $\sigma_{1.234 \ldots n}$ as follows

$$\sigma_{(\text{est.}\, X_1)} = \sigma_{1.234 \ldots n} \qquad (96)$$

(standard error of estimate for n *variables)*

Since $\sigma_{1.234\ldots n}$ must be computed in order to evaluate the partial regression coefficients (p. 421), $\sigma_{(\text{est.}\,X_1)}$ is always calculated in the course of the problem. In Table 59, the $\sigma_{(\text{est.}\,X_1)}$ of a prediction of honor points is 6.3. The chances are about seven in ten or two in three, that the "most probable" honor point score forecast for *any* student will be in error by six points or *less*.

It is worth while examining further into the meaning of $\sigma_{(\text{est.}\,X_1)}$. This standard error of estimate equals $\sigma_{1.23}$; and the latter indicates the effect upon the variability of Test 1 (honor points) obtained by eliminating (or holding constant) the influence of Tests 2 and 3 (general intelligence and study effort). The *smaller* $\sigma_{1.23}$ is with respect to σ_1, the greater the influence exerted by our two factors upon Test 1's variability. In Table 59 it is clear that in ruling out the variability in Test 1 attributable to Tests 2 and 3, we reduce σ_1 from 11.2 to 6.3 ($\sigma_{1.23}$) or by nearly one-half. This means that students *alike* in general intelligence and in study habits differ much *less* in scholastic achievement than do students in general.

From the multiple regression equation $\overline{X}_1 = .57X_2 + 1.12X_3 - 66$ (see p. 413), X_1 (honor points) can be predicted with a *smaller error of estimate* than from any other *linear* equation. Put differently, the standard error of estimate is a *minimum* when the regression equation is used to estimate X_1 scores.* Hence, the values of X_1 predicted from the multiple regression equation are the "best estimates" of the actual X_1 values which can be made from a linear equation containing the given variables.

5. The Coefficient of Multiple Correlation, R

(1) General Formulas

The correlation between a single dependent or criterion variable X_1 and $(n - 1)$ independent variables combined by means of a multiple regression equation is given by the formula

* Yule, G. U., and Kendall, M. C., *An Introduction to the Theory of Statistics* (12 ed., 1940), pp. 262–267.

$$R_{1(23\ldots n)} = \sqrt{1 - \frac{\sigma^2_{1.23\ldots n}}{\sigma^2_1}} \tag{97}$$

(*multiple correlation coefficient in terms of partial σ's* — n *variables*)

in which

$R_{1(23\ldots n)}$ = the coefficient of multiple correlation

σ_1 = the standard deviation of the criterion (X_1) scores

$\sigma_{1.23\ldots n}$ = the variability left in Test 1 when the variability of Tests 2, 3 . . . n is held constant through partial correlation.

When there are only three variables, the multiple coefficient of correlation becomes

$$R_{1(23)} = \sqrt{1 - \frac{\sigma^2_{1.23}}{\sigma^2_1}}$$

when there are five variables

$$R_{1(2345)} = \sqrt{1 - \frac{\sigma^2_{1.2345}}{\sigma^2_1}}$$

If we replace $\sigma_{1.23\ldots n}$ in formula (97) by its value in terms of the entire and partial r's [see formula (88)] we may write the general formula for $R_{1(234\ldots n)}$ as follows:

$$R_{1(234\ldots n)} = \sqrt{1 - [(1 - r^2_{12})(1 - r^2_{13.2})\ldots(1 - r^2_{1n.23\ldots(n-1)})]} \tag{98}$$

(*multiple coefficient of correlation in terms of partial coefficients of correlation* — n *variables*)

Since a higher order σ may be written in a variety of ways, the number depending upon its order (see p. 417), there are several alternate forms for R. These serve as valuable means of checking the accuracy of our arithmetical calculations. In a three-variable problem, for example, $R_{1(23)}$ may be written as

$$R_{1(23)} = \sqrt{1 - [(1 - r^2_{12})(1 - r^2_{13.2})]}$$

or as

$$R_{1(32)} = \sqrt{1 - [(1 - r^2_{13})(1 - r^2_{12.3})]}$$

The standard error of estimate is a minimum when the multiple regression equation is employed in estimating X_1 scores (p. 423). Hence the multiple coefficient of correlation, R, is the *maximum correlation* obtainable between actual X_1 scores and $\bar{X}_1$ scores estimated from a knowledge of the variables X_2, X_3 . . . X_n in the regression equation. The truth of this statement is contingent upon linearity of regression in *all* of the correlations. R indicates how accurately a given combination of variables represents the actual values of X_1 (the criterion) when our test scores are combined in accordance with the "best" linear equation.

(2) Multiple R in Terms of β Coefficients

R^2 may be expressed in terms of the beta coefficients and the zero order r's:

$$R^2_{1(23\ldots n)} = \beta_{12.34\ldots n}r_{12} + \beta_{13.24\ldots n}r_{13} + \cdots + \beta_{1n.23\ldots(n-1)}r_{1n} \tag{99}$$

(*multiple* R² *in terms of* β *coefficients and zero order* r'*s*)

For three variables (99) becomes

$$R^2_{1(23)} = \beta_{12.3}r_{12} + \beta_{13.2}r_{13}$$

From page 422 we find $\beta_{12.3} = .81$ and $\beta_{13.2} = .60$; and from Table 59 that $r_{12} = .60$ and $r_{13} = .32$. Substituting in (99) above, we get

$$\begin{aligned} R^2_{1(23)} &= .81 \times .60 + .60 \times .32 \\ &= .49 + .19 \\ R^2_{1(23)} &= .68 \\ R_{1(23)} &= .83 \end{aligned}$$

$R^2_{1(23\ldots n)}$ gives the proportion of the variance of the criterion measure (X_1) attributable to the joint action of the variables X_2, $X_3 \ldots X_n$. As shown above, $R^2_{1(23)} = .68$; and, accordingly, 68% of whatever makes freshmen differ in (1) school achievement, can be attributed to differences in (2) general intelligence, and (3) study habits. By means of formula (99) the total contribution of .68 can be broken down further into the

independent contributions of general intelligence (X_2) and study habits (X_3). Thus from the equation $R^2_{1(23)} = .49 + .19$, we know that 49% is the contribution of general intelligence to the variance of honor points, and 19% is the contribution of study habits. The remaining 32% of the variance of X_1 must be attributed to factors not measured in our problem.

(3) The Significance of R

Multiple R is positive,* always less than 1.00, and always greater than the correlation coefficients r_{12}, r_{13}, . . . r_{1n}. The significance of an R can best be tested, perhaps, against the null hypothesis by means of Table 61. This table must be entered

TABLE 61

COEFFICIENTS OF CORRELATION SIGNIFICANT AT THE 5% LEVEL AND AT THE 1% LEVEL FOR VARYING DEGREES OF FREEDOM

Degrees of Freedom	Number of Variables						
	2	3	4	5	6	7	9
1	.997 **1.000**	.999 **1.000**	.999 **1.000**	.999 **1.000**	1.000 **1.000**	1.000 **1.000**	1.000 **1.000**
2	.950 **.990**	.975 **.995**	.983 **.997**	.987 **.998**	.990 **.998**	.992 **.998**	.994 **.999**
3	.878 **.959**	.930 **.976**	.950 **.983**	.961 **.987**	.968 **.990**	.973 **.991**	.979 **.993**
4	.811 **.917**	.881 **.949**	.912 **.962**	.930 **.970**	.942 **.975**	.950 **.979**	.961 **.984**
5	.754 **.874**	.836 **.917**	.874 **.937**	.898 **.949**	.914 **.957**	.925 **.963**	.941 **.971**
6	.707 **.834**	.795 **.886**	.839 **.911**	.867 **.927**	.886 **.938**	.900 **.946**	.920 **.957**
7	.666 **.798**	.758 **.855**	.807 **.885**	.838 **.904**	.860 **.918**	.876 **.928**	.900 **.942**
8	.632 **.765**	.726 **.827**	.777 **.860**	.811 **.882**	.835 **.898**	.854 **.909**	.880 **.926**
9	.602 **.735**	.697 **.800**	.750 **.836**	.786 **.861**	.812 **.878**	.832 **.891**	.861 **.911**

* Since R is always positive, chance errors are cumulative and may be considerable if the sample is small and the number of variables large. For the correction of R for chance errors, see Formula 100, page 451.

TABLE 61—*Continued*

Degrees of Freedom	Number of Variables						
	2	3	4	5	6	7	9
10	.576	.671	.726	.763	.790	.812	.843
	.708	**.776**	**.814**	**.840**	**.859**	**.874**	**.895**
11	.553	.648	.703	.741	.770	.792	.826
	.684	**.753**	**.793**	**.821**	**.841**	**.857**	**.880**
12	.532	.627	.683	.722	.751	.774	.809
	.661	**.732**	**.773**	**.802**	**.824**	**.841**	**.866**
13	.514	.608	.664	.703	.733	.757	.794
	.641	**.712**	**.755**	**.785**	**.807**	**.825**	**.852**
14	.497	.590	.646	.686	.717	.741	.779
	.623	**.694**	**.737**	**.768**	**.792**	**.810**	**.838**
15	.482	.574	.630	.670	.701	.726	.765
	.606	**.677**	**.721**	**.752**	**.776**	**.796**	**.825**
16	.468	.559	.615	.655	.686	.712	.751
	.590	**.662**	**.706**	**.738**	**.762**	**.782**	**.813**
17	.456	.545	.601	.641	.673	.698	.738
	.575	**.647**	**.691**	**.724**	**.749**	**.769**	**.800**
18	.444	.532	.587	.628	.660	.686	.726
	.561	**.633**	**.678**	**.710**	**.736**	**.756**	**.789**
19	.433	.520	.575	.615	.647	.674	.714
	.549	**.620**	**.665**	**.698**	**.723**	**.744**	**.778**
20	.423	.509	.563	.604	.636	.662	.703
	.537	**.608**	**.652**	**.685**	**.712**	**.733**	**.767**
21	.413	.498	.552	.592	.624	.651	.693
	.526	**.596**	**.641**	**.674**	**.700**	**.722**	**.756**
22	.404	.488	.542	.582	.614	.640	.682
	.515	**.585**	**.630**	**.663**	**.690**	**.712**	**.746**
23	.396	.479	.532	.572	.604	.630	.673
	.505	**.574**	**.619**	**.652**	**.679**	**.701**	**.736**
24	.388	.470	.523	.562	.594	.621	.663
	.496	**.565**	**.609**	**.642**	**.669**	**.692**	**.727**
25	.381	.462	.514	.553	.585	.612	.654
	.487	**.555**	**.600**	**.633**	**.660**	**.682**	**.718**
26	.374	.454	.506	.545	.576	.603	.645
	.478	**.546**	**.590**	**.624**	**.651**	**.673**	**.709**
27	.367	.446	.498	.536	.568	.594	.637
	.470	**.538**	**.582**	**.615**	**.642**	**.664**	**.701**
28	.361	.439	.490	.529	.560	.586	.629
	.463	**.530**	**.573**	**.606**	**.634**	**.656**	**.692**

TABLE 61—*Continued*

Degrees of Freedom	Number of Variables						
	2	3	4	5	6	7	9
29	.355	.432	.482	.521	.552	.579	.621
	.456	**.522**	**.565**	**.598**	**.625**	**.648**	**.685**
30	.349	.426	.476	.514	.545	.571	.614
	.449	**.514**	**.558**	**.591**	**.618**	**.640**	**.677**
35	.325	.397	.445	.482	.512	.538	.580
	.418	**.481**	**.523**	**.556**	**.582**	**.605**	**.642**
40	.304	.373	.419	.455	.484	.509	.551
	.393	**.454**	**.494**	**.526**	**.552**	**.575**	**.612**
45	.288	.353	.397	.432	.460	.485	.526
	.372	**.430**	**.470**	**.501**	**.527**	**.549**	**.586**
50	.273	.336	.379	.412	.440	.464	.504
	.354	**.410**	**.449**	**479**	**.504**	**.526**	**.562**
60	.250	.308	.348	.380	.406	.429	.467
	.325	**.377**	**.414**	**.442**	**.466**	**.488**	**.523**
70	.232	.286	.324	.354	.379	.401	.438
	.302	**.351**	**.386**	**.413**	**.436**	**.456**	**.491**
80	.217	.269	.304	.332	.356	.377	.413
	.283	**.330**	**.362**	**.389**	**.411**	**.431**	**.464**
90	.205	.254	.288	.315	.338	.358	.392
	.267	**.312**	**.343**	**.368**	**.390**	**.409**	**.441**
100	.195	.241	.274	.300	.322	.341	.374
	.254	**.297**	**.327**	**.351**	**.372**	**.390**	**.421**
125	.174	.216	.246	.269	.290	.307	.338
	.228	**.266**	**.294**	**.316**	**.335**	**.352**	**.381**
150	.159	.198	.225	.247	.266	.282	.310
	.208	**.244**	**.270**	**.290**	**.308**	**.324**	**.351**
200	.138	.172	.196	.215	.231	.246	.271
	.181	**.212**	**.234**	**.253**	**.269**	**.283**	**.307**
300	.113	.141	.160	.176	.190	.202	.223
	.148	**.174**	**.192**	**.208**	**.221**	**.233**	**.253**
400	.098	.122	.139	.153	.165	.176	.194
	.128	**.151**	**.167**	**.180**	**.192**	**.202**	**.220**
500	.088	.109	.124	.137	.148	.157	.174
	.115	**.135**	**.150**	**.162**	**.172**	**.182**	**.198**
1000	.062	.077	.088	.097	.105	.112	.124
	.081	**.096**	**.106**	**.115**	**.122**	**.129**	**.141**

with $N - m$ degrees of freedom, and with the number of variables (m) in the problem. To illustrate with Table 59, $R = .83$, $N = 450$, $m = 3$ and $N - m = 450 - 3$ or 447. From the column headed "3" in Table 61 we read that for 447 degrees of freedom the R's at the .05 and .01 levels (by interpolation) are .116 and .143. Only once in twenty trials would an R of .116 arise by sampling fluctuations on the null hypothesis, and only once in 100 trials would an R of .143 occur. As our R is very much larger than .14, it is highly significant. Table 61 may be used with problems involving up to nine variables. Suppose that $R_{1(2345)} = .526$ and $N = 40$. From the column headed "5 variables" in Table 61, we find that for $40 - 5$ or 35 degrees of freedom, the R's are .482 and .556 at the .05 and .01 levels. The obtained R is significant, therefore, at the .05, but not at the .01, level.

IV. Spurious Correlation

The correlation between two sets of test scores is said to be *spurious* when it is due in some part, at least, to factors other than those which determine performance in the tests themselves. In general, the cause of spurious correlation lies in a failure to control conditions; and the most usual effect of this lack of control is a "boosting" or inflation of the coefficient. Some of the situations which may lead to spurious correlation will be given in this section.

1. Spurious Correlation Arising from Heterogeneity

We have shown elsewhere (p. 404) how a lack of uniformity in age conditions will lead to correlations which are spuriously high. Failure to take account of heterogeneity introduced by the age factor is a prolific source of error in correlational work. To cite an example, within a group of boys ten to eighteen years old, a substantial correlation will appear between strength of grip and memory span, quite apart from any intrinsic relationship, due solely to the fact that both variables increase with age. In stating the correlation between two tests, or the

reliability coefficient of a test, one should always be careful to specify the range of ages, grades included, and other data bearing upon physical, mental, and cultural differences, in order to show the degree of heterogeneity in the group. Without this information, the r may be of little value.

Heterogeneity is introduced by other factors than age. If alcoholism, degeneracy, and bad heredity are all positively related, the r between alcoholism and degeneracy will be too high (because of the effect of heredity upon both factors) unless heredity can be "held constant." Again, assume that we have measured two distinctly different groups, 500 college seniors, and 500 day laborers, upon a cancellation test and upon a general intelligence test. The mean ability in both tests will be definitely higher in the college group. Now even if the correlation between the two tests is zero within each group taken separately, if the two groups are combined a positive correlation will appear because of the heterogeneity of the group with respect to age, intelligence, and educational background. Such a correlation is, of course, spurious.*

To be a valid measure of relationship, a correlation coefficient must be freed of the extraneous influences which affect the relationship between the variables concerned. This may be accomplished (1) by selecting samples or groups in which age (or whatever the factor to be controlled) is constant; or (2) one may use partial correlation if the factor to be controlled can be measured and its correlation with the variables studied can be calculated.

2. Spurious Index Correlation †

Even when three variables X_1, X_2, and X_3 are uncorrelated, a correlation between the indices Z_1 and Z_2 (where $Z_1 = X_1/X_3$,

* Garrett, H. E., and Anastasi, A., *The Tetrad-Difference Criterion and the Measurement of Mental Traits*, Annals New York Academy of Sciences, No. 33 (1932), 233–282.

† Yule, G. U., *An Introduction to the Theory of Statistics* (1932), pp. 215–216.

Thomson, G. H., and Pintner, R., "Spurious Correlation and Relationship Between Tests," *Journal of Educational Psychology*, 15 (1924), 433–444.

and $Z_2 = X_2/X_3$) may appear which is as large as .50. To illustrate, if two individuals observe a series of magnitudes (e.g., Galton bar settings) independently, the absolute errors of observation (X_1 and X_2) may be uncorrelated, and still an appreciable correlation appear between the errors made by the two observers, when these are expressed as *percents* of the observed magnitudes (X_3). The spurious element here, of course, is the common factor X_3 in the denominator of the ratios.

One of the commonest examples of a spurious index relationship in psychology is found in the correlation of I.Q.'s or E.Q.'s obtained from intelligence and achievement tests. If the I.Q.'s of 500 children ranging in age from three to fourteen years are calculated from two tests X_1 and X_2, the correlation is between $\frac{\text{M.A.}_1}{\text{C.A.}}$ and $\frac{\text{M.A.}_2}{\text{C.A.}}$. If C.A. were a *constant* (the same for all children) it would have no effect on the correlation and we would simply be correlating M.A.'s. But when C.A. varies from child to child there is usually a correlation between C.A. and M.A. which tends to increase the r between I.Q.'s — sometimes considerably.

3. Spurious Correlation between Averages

Spurious correlation usually results when the average scores made by a number of different groups on a given test are correlated against the average scores made by the same groups on a second test. An example is furnished by the correlations reported by Bagley* between the *mean* Army Alpha scores, by states, and such "educational" factors as number of schools, books sold, magazines circulated in the states, etc. Most of these correlations are high — many above .90. If average correlations by states are compared with the correlations between intelligence scores and number of years spent in school within the separate states, these latter r's are usually much lower. Correlations between averages become "inflated" because a large number of factors which ordinarily reduce the correlation

* Bagley, W. C., *Determinism in Education* (1925), p. 81.

within a single group cancel out when averages are taken from group to group. Average intelligence test scores, for instance, increase regularly as we go up the occupational scale from day laborer to the professions; but the correlation between intelligence and status (training, salary, etc.) at a given occupational level is far from perfect.

PROBLEMS

1. The correlation between a general intelligence test and school achievement in a group of children from eight to fourteen years old is .80. The correlation between the general intelligence test and age in the same group is .70; and the correlation between school achievement and age is .60. What is the correlation between general intelligence and school achievement in children of the same age? Comment upon your result.
2. In a group of 100 college freshmen, the correlation between (1) Army Alpha and (2) the A-cancellation test is .20. The correlation between (1) Army Alpha and (3) a battery of controlled association tests in the same group is .70. If the correlation between (2) cancellation and (3) controlled association is .45, what is the "net" correlation between Army Alpha and cancellation in this group? Between Alpha and controlled association? Interpret your results.
3. Explain why some variables are of such a nature that it is difficult to hold them "constant," and hence to employ them in problems involving partial correlation.
4. Given the following data for fifty-six children:

$$X_1 = \text{Stanford-Binet I.Q.}$$
$$X_2 = \text{Memory for Objects}$$
$$X_3 = \text{Cube Imitation}$$

$M_1 = 101.71$	$M_2 = 10.06$	$M_3 = 3.35$
$\sigma_1 = 13.65$	$\sigma_2 = 3.06$	$\sigma_3 = 2.02$
$r_{12} = .41$	$r_{13} = .50$	$r_{23} = .16$

(a) Work out the regression equation of X_2 and X_3 upon X_1, using the method of Section II.

(b) Compute $R_{1(23)}$ and $\sigma_{(\text{est. } X_1)}$.

(c) If a child's score is 12 in Test X_2 and 4 in Test X_3, what is his most probable score in X_1 (I.Q.)?

5. Let X_1 be a criterion and X_2 and X_3 be two other tests. Correlations and σ's are as follows:

$r_{12} = .60$	$\sigma_1 = 5.00$
$r_{13} = .50$	$\sigma_2 = 10.00$
$r_{23} = .20$	$\sigma_3 = 8.00$

How much more accurately can X_1 be predicted from X_2 and X_3 than from either alone?

6. Given a team of two tests, each of which correlates .50 with a criterion. If the two tests correlate .20
   (*a*) How much would the addition of another test which correlates .50 with the criterion and .20 with each of the other tests improve the predictive value of the team?
   (*b*) How much would the addition of two such tests improve the predictive value of the team?

7. Two absolutely independent tests B and C completely determine the criterion A. If B correlates .50 with A, what is the correlation of C and A? What is the multiple correlation of A with B and C?

8. Comment upon the following statements:
   (*a*) It is good practice to correlate E.Q.'s achieved upon two educational achievement tests, no matter how wide the age range.
   (*b*) The positive correlation between average Army Alpha scores by states and the average elevation of the states above sea level proves the close relationship of intelligence and geography.
   (*c*) The correlation between memory test scores and tapping rate in a group of 200 eight-year-old children is .20; and the correlation between memory test scores and tapping rate in a group of 100 college freshmen is .10. When the two groups are combined the correlation between these two tests becomes .40. This shows that we must have large groups in order to get high correlations.

Answers

1. $r = .67$
2. r (Alpha and cancellation) $= -.19$; r (Alpha and controlled association) $= .70$
4. (*a*) $\bar{X}_1 = 1.47X_2 + 2.98\,X_3 + 76.95$
   (*b*) $R_{1(23)} = .60$; $\sigma_{(\text{est. } X_1)} = 10.93$
   (*c*) 106.50 or 107

5. From X_2 alone, $\sigma_{(\text{est. } X_1)} = 4.0$
   From X_3 alone, $\sigma_{(\text{est. } X_1)} = 4.3$
   From X_2 and X_3, $\sigma_{(\text{est. } X_1)} = 3.5$
6. (*a*) R increases from .64 to .73
   (*b*) R increases from .64 to .79
7. $r_{AC} = .87$; $R_{A(BC)} = 1.00$

CHAPTER XIV

MULTIPLE CORRELATION IN TEST SELECTION

I. The Wherry-Doolittle Test Selection Method*

The method of solving multiple correlation problems outlined in Section II and Table 59 of Chapter XIII is adequate enough when there are only three (or not more than four) variables. In problems involving more than four variables, however, the mechanics of calculation become almost prohibitive unless some systematic scheme of solution is adopted. The Wherry-Doolittle Test Selection Method, to be presented in this section, provides a method of solving multiple correlation problems with a minimum of statistical labor. This method selects the tests of the battery analytically and adds them one at a time until a maximum R is obtained. To illustrate, suppose we wish to predict aptitude for a certain technical job in a factory. Criterion ratings for job proficiency have been obtained and eight tests tried out as possible indicators of job aptitude. By use of the Wherry-Doolittle method we can (1) select those tests (e.g., three or four) which yield a maximum R with the criterion and discard the rest; (2) calculate the multiple R after the addition of each test, stopping the process when R no longer increases; (3) compute a multiple regression equation from which the criterion can be predicted with the highest precision of which the given list of tests is capable.

The application of the Wherry-Doolittle test selection method to an actual problem is shown in Example (1) below. Steps in computation are outlined in order and are illustrated by reference to the data of Example (1), so that the reader may follow the process in detail.

* Stead, W. H., Shartle, C. L., et al., *Occupational Counseling Techniques* (1940), Appendix 5

1. Solution of a Multiple Correlation Problem by the Wherry-Doolittle Test Selection Method

Example (1) In Table 62 are presented the intercorrelations of ten tests administered in the Minnesota study of Mechanical Ability. The criterion — called the "quality" criterion — was a measure of the excellence of mechanical work done by 100 junior high-school boys. The tests in Table 62 are fairly representative of the wide range of measures used in the Minnesota study. Our immediate problem is to choose from among these variables the most valid battery of tests, i.e., those tests which will predict the criterion most efficiently. Selection of tests is made by the Wherry-Doolittle method.

TABLE 62

INTERCORRELATIONS OF TEN TESTS AND A CRITERION

(Data from the Minnesota Study of Mechanical Ability *)

List of Tests ($N = 100$)

C = Quality criterion
1 = Packing blocks
2 = Card sorting
3 = Minnesota spatial relations boards, A, B, C, D
4 = Paper form boards, A and B
5 = Stenquist Picture I
6 = Stenquist Picture II
7 = Minnesota assembly boxes, A, B, C
8 = Mechanical operations questionnaire
9 = Interest analysis blank
10 = Otis intelligence test

	1	2	3	4	5	6	7	8	9	10
C	.26	.19	.53	.52	.24	.31	.55	.30	.55	.26
1		.52	.34	.14	.18	.21	.30	.00	.34	.00
2			.23	.14	.10	.24	.13	−.12	.23	.08
3				.63	.42	.39	.56	.22	.55	.23
4					.37	.30	.49	.24	.61	.56
5						.54	.46	.24	.23	.11
6							.40	.19	.13	.21
7								.40	.41	.13
8									.25	.18
9										.38

Steps in the solution of Example (1) may be outlined in order.

* Paterson, D. G., Elliott, R. M., et al., *Minnesota Mechanical Ability Tests* (1930), Appendix 4.

Step 1

Draw up work sheets like those of Tables 63 and 64. The correlation coefficients between tests and criterion are entered in Table 62.

Step 2

Enter these coefficients *with signs reversed* in the V_1 row of Table 63.* The numbers heading the columns refer to the tests.

TABLE 63

Tests

	1	2	3	4	5	6	7	8	9	10
V_1	−.260	−.190	−.530	−.520	−.240	−.310	−.550	−.300	−.550	−.260
V_2	−.095	−.118	−.222	−.250	.013	−.090		−.080	−.324	−.188
V_3	−.010	−.049	−.097	−.091	.029	−.077		−.047		−.061
V_4	.005	−.034		−.057	.004	−.046		−.053		−.056
V_5	−.012	−.039			.012	−.039		−.051		−.018

$$\frac{V_1^2}{Z_1} = \frac{(-.550)^2}{1.000} = .3025; \quad \frac{V_2^2}{Z_2} = \frac{(.324)^2}{.832} = .1261; \quad \frac{V_3^2}{Z_3} = \frac{(.097)^2}{.563} = .0167; \quad \frac{V_4^2}{Z_4} = \frac{(.057)^2}{.489} = .0066; \quad \frac{V_5^2}{Z_5} = \frac{(-.051)^2}{.829} = .0031$$

Step 3

Enter the numbers 1.000 in each column of the row Z_1 in Table 64.

TABLE 64

Tests

	1	2	3	4	5	6	7	8	9	10
Z_1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Z_2	.910	.983	.686	.760	.788	.840		.840	.832	.983
Z_3	.853	.945	.563	.559	.786	.839		.831		.854
Z_4	.839	.931		.489	.748	.782		.829		.852
Z_5	.796	.927			.737	.775		.829		.637

$$\frac{1}{.832} = 1.202$$

$$\frac{1}{.563} = 1.776$$

$$\frac{1}{.489} = 2.045$$

* Correlation coefficients are assumed to be accurate to three or to four decimals in subsequent calculations to avoid the loss of precision which results when decimals are rounded to two places. (See p. 455.)

Step 4

Select that test having the highest $\frac{V_1^2}{Z_1}$ quotient as the *first* test of the battery. From Tables 63 and 64 we find that Tests 7 and 9 both have correlations of .550 with the criterion, and that these are the largest r's in the table. Either Test 7 or Test 9 could be selected as the first test of our battery. We have chosen Test 7 because it is the more objective measure of performance.

Step 5

Apply the Wherry shrinkage formula

$$\bar{R}^2 = 1 - K^2\left(\frac{N-1}{N-m}\right)$$

in which $\bar{R}$ is the "shrunken" multiple correlation coefficient, the coefficient from which chance error has been removed.* This corrected R may be calculated in a systematic way as follows:

(1) Prepare a work sheet similar to that shown in Table 65.

TABLE 65

a	b	c	d	e	f	g	
	$\frac{V_m^2}{Z_m}$	K^2	$\frac{N-1}{N-m}$	$\bar{K}^2$	$\bar{R}^2$	$\bar{R}$	Test #
0		1.000	$(N-100)$				
1	.3025	.6975	1.000	.6975	.3025	.5500	7
2	.1261	.5714	1.010	.5771	.4229	.6503	9
3	.0167	.5547	1.021	.5663	.4337	.6586	3
4	.0066	.5481	1.031	.5651	.4349	.6595	4
5	.0031	.5450	1.042	.5679	.4321	.6573	8

(2) Enter 1.000 in column c, row 0, under K^2. Enter $N = 100$ in column d.

* Wherry, R. J., *A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation*, Annals of Mathematical Statistics, Vol. 2 (1931), 440–451.

(3) Enter the quotient $\frac{V_1^2}{Z_1}$ in column b, row 1. $\frac{V_1^2}{Z_1} = \frac{(-.550)^2}{1.000} = .3025$ *

(4) Subtract .3025 from 1.000 to give .6975 as the entry in column c under K^2.

(5) Find the quotient $\frac{(N-1)}{(N-m)}$ and record it in column d. $(N-1) = 99$; and since m (number of tests selected) is 1, $(N-m)$ also equals 99 and $\frac{(N-1)}{(N-m)} = 1.0000$.

(6) Write the product of columns c and d in column e: $.6975 \times 1.000 = .6975$.

(7) Subtract the column e entry from 1.0000 to obtain $\bar{R}^2$ (the shrunken multiple correlation coefficient) in column f. In Table 65 the $\bar{R}^2$ entry, of course, is .3025.

(8) Find the square root of the column f entry and enter the result in column g under $\bar{R}$. Our entry is .5500, the correlation of Test 7 with the criterion. No correction for chance errors is necessary for one test.

Step 6

To aid in the selection of a *second* test to be added to our battery of one, a work sheet similar to that shown in Table 66 should be prepared. Calculations in Table 66 are as follows:

(1) Leave a_1 row blank.

(2) Enter in row b_1 the correlations of Test 7 (*first* selected test) with each of the other tests in Table 62. These r's are .300, .130, .560, etc., and are entered in the columns numbered to correspond to the tests. Enter 1.000 in the column for Test 7. In column $-$ C enter the correlation of Test 7 with the criterion *with sign reversed,* i.e., as $-.550$.

(3) Write the algebraic sum of the b_1 entries in the "Check Sum" column. This sum is 3.730.

* Quotient is taken to four decimals (p. 455).

TABLE 66

	1	2	3	4	5	6	7	8	9	10	− C	Check Sum	Test #
a_1	—	—	—	—	—	—	—	—	—	—	—	—	
b_1	.300	.130	.560	.490	.460	.400	1.000	.400	.410	.130	−.550	3.730	.7
c_1	−.300	−.130	−.560	−.490	−.460	−.400	−1.000	−.400	−.410	−.130	.550	−3.730	
a_2	.340	.230	.550	.610	.230	.130	.410	.250	1.000	.380	−.550	3.580	9
b_2	.217	.177	.320	.409	.041	−.034	|	.086	.832	.327	−.324	2.051	
c_2	−.261	−.213	−.385	−.492	−.049	.041	|	−.103	−1.000	−.393	.389	−2.465	
a_3	.340	.230	1.000	.630	.420	.390	.560	.220	.550	.230	−.530	4.040	3
b_3	.088	.089	.563	.199	.146	.179	|	−.037	|	.031	−.097	1.161	
c_3	−.156	−.158	−1.000	−.353	−.259	−.318	|	.066	|	−.055	.172	−2.062	
a_4	.140	.140	.630	1.000	.370	.300	.490	.240	.610	.560	−.520	3.960	4
b_4	−.145	−.042	|	.489	.073	.058	|	.015	|	.324	−.057	.715	
c_4	.297	.086	|	−1.000	−.149	−.119	|	−.031	|	−.663	.117	−1.462	

(4) Multiply each b_1 entry by the *negative reciprocal* of the b_1 entry for Test 7, the *first* selected test. Enter these products in the c_1 row. Since the negative reciprocal of Test 7's b_1 entry is -1.000, we need simply write the b_1 entries in the c_1 row with signs reversed.

Step 7

Draw a vertical line under Test 7 in Table 63 to show that it has been selected. To select a *second* test proceed as follows:

(1) To each V_1 entry in Table 63, add algebraically the product of the b_1 entry in the criterion ($-$ C) column of Table 66 by the c_1 entry for each of the other tests. Enter results in the V_2 row. The formula for V_2 is $V_2 = V_1 + b_1$ (criterion) $\times$ c_1 (each test). To illustrate, from Table 66 and Table 63 we have

For Test 1: $V_2 = -.260 + (-.550) \times (-.300) = -.260 + .165 = -.095$

For Test 4: $V_2 = -.520 + (-.550) \times (-.490) = -.520 + .270 = -.250$

For Test 9: $V_2 = -.550 + (-.550) \times (-.410) = -.550 + .226 = -.324$

(2) To each Z_1 in Table 64 add algebraically the product of the b_1 and c_1 entries for each test got from Table 66. Enter these results in the Z_2 row. The formula is $Z_2 = Z_1 + b_1$ (a given test) $\times$ c_1 (same test). To illustrate, from Tables 63 and 66

For Test 1: $Z_2 = 1.000 + (.300) \times (-.300) = 1.000 - .090 = .910$

For Test 4: $Z_2 = 1.000 + (.490) \times (-.490) = 1.000 - .240 = .760$

For Test 9: $Z_2 = 1.000 + (.410) \times (-.410) = 1.000 - .168 = .832$

Step 8

Now select the test having the largest $\frac{V^2_2}{Z_2}$ quotient, as the

second test for our battery. The quantity $\frac{V_2^2}{Z_2}$ is a measure of the amount which the second test contributes to the squared multiple correlation coefficient, $\bar{R}^2$. From Tables 63 and 64 we find that Test 9 has the largest $\frac{V_2^2}{Z_2}$ quotient: $\frac{(.324)^2}{.832} = .1261$.

Step 9

To calculate the new multiple correlation coefficient when Test 9 is added to Test 7, proceed as follows:

(1) The quantity .1261 $\left(\frac{V_2^2}{Z_2}\right)$ is entered in column b, row 2 of Table 65.

(2) Subtract the ratio $\frac{V_2^2}{Z_2}$ from the K^2 entry in column c, row 1, and enter the result in column c, row 2; e.g., for the entry in column c, row 2, we have $.6975 - .1261$, or .5714.

(3) Find the quotient $\frac{(N-1)}{(N-m)}$. Since $N = 100$ and m (number of tests chosen) = 2, we have $\frac{(N-1)}{(N-m)}$ or $\frac{99}{98} = 1.010$, as the column d, row 2 entry.

(4) Record the product of the c and d columns in column e: $.5714 \times 1.010 = .5771$.

(5) Subtract .5771 (column e) from 1.0000 to give .4229 as the entry in column f, row 2.

(6) Take the square root of .4229 and enter the result, .6503, in column g. This is the multiple coefficient $\bar{R}$ corrected for chance errors. It is clear that by adding Test 9 to Test 7 we increase $\bar{R}$ from .5500 to .6503, a substantial gain.

Step 10

Since $\bar{R}$ for Tests 7 and 9 is larger than the correlation for Test 7 alone, we proceed to add a *third* test in the hope of further increasing the multiple $\bar{R}$. The procedure is shown in Step 11.

Step 11

Return to Table 66 and

(1) Record in the a_2 row the correlation coefficient of the *second* selected test (i.e., Test 9) with each of the other tests *and* with the criterion. (Read r's from Table 62.) The correlation of Test 9 with the criterion is entered *with sign reversed* (i.e., as $-.550$).

(2) Enter the algebraic sum of the a_2 entries (i.e., 3.580) in the Check Sum column.

(3) Draw a vertical line down through the b_2 and c_2 rows for Test 7, the *first* selected test. This indicates that Test 7 has already been chosen.

(4) Compute the b_2 entry for each test by adding to the a_2 entry the product of the b_1 entry of the given test by the c_1 entry of the *second* selected test (i.e., Test 9). The formula is $b_2 = a_2 + b_1$ (given test) $\times$ c_1 (*second* selected test). To illustrate:

For Test 2: $b_2 = .230 + (.130)(-.410) = .230 - .053 = .177$

For Test 6: $b_2 = .130 + (.400)(-.410) = .130 - .164 = -.034$

For Test 10: $b_2 = .380 + (.130)(-.410) = .380 - .053 = .327$

Compute b_2 entries for criterion and Check Sum column in the same way. For the criterion column we have $-.550 + (-.550)(-.410)$ or $-.324$. For the Check Sum column we have $3.580 + (3.730)(-.410)$ or 2.051.

(5) There are three checks for the b_2 row. (*a*) The entry for the *second* selected test (Test 9) should equal the Z_2 entry for the same test in Table 64. Note that both entries are .832. (*b*) The entry in the criterion column should equal the V_2 entry of the second selected test (Test 9) in Table 63; both entries are $-.324$. (*c*) The entry in the Check Sum column should equal the sum of all of the

entries in the b_2 row. Adding .217, .177, .320, etc., we get 2.051, checking our calculations to the third decimal.

(6) Multiply each b_2 entry by the *negative reciprocal* of the b_2 entry for the *second* selected test (Test 9), and record results in the c_2 row. The negative reciprocal of .832 is -1.202. The c_2 entry for Test 1 is $.217 \times -1.202$ or $-.261$; for Test 2, $-.177 \times -1.202$ or $-.213$; and so on for the other tests. For the criterion column the c_2 entry is $(-.324) \times -1.202$ or .389; and for the Check Sum the c_2 entry is 2.051×-1.202 or -2.465.

(7) There are three checks for the c_2 entries. (*a*) The c_2 row entry of the second selected test (Test 9) should be -1.000. (*b*) The c_2 entry in the Check Sum column should equal the sum of all c_2 entries. Adding the c_2 entries in Table 66, we find the sum to be -2.465, the Check Sum entry. (*c*) The product of the b_2 and c_2 entries in the criterion column should equal the quotient $\frac{V_2^2}{Z_2}$ in column b, row 2 of Table 65 in absolute value. Note that the product $(-.324 \times .389) = -.1261$, thus checking our entry (disregard signs).

Step 12

Draw a vertical line under Test 9 in Table 63, to indicate that it has been selected as our second test. Then proceed as in Step 7 to compute V_3 and Z_3 in order to select a *third* test. The formula for V_3 is $V_3 = V_2 + b_2$ (criterion) $\times\, c_2$ (each test). The formula for Z_3 is $Z_3 = Z_2 + b_2$ (a given test) $\times\, c_2$ (same test). The third selected test is that one which has the largest $\frac{V_3^2}{Z_3}$ quotient in Table 63. This is Test 3, for which $V_3 = -.222 + (-.324)(-.385)$ or $-.097$; and $Z_3 = .686 + (.320)(-.385) = .563$. The quotient $\frac{V_3^2}{Z_3} = .0167$.

Step 13

Entering .0167 $\left(\frac{V_3^2}{Z_3}\right)$ in column b, row 3, of Table 65, follow

the procedure of Step 9 to get $\overline{R} = .6586$. Note that $\frac{(N-1)}{(N-m)}$ $= 99/97$ or 1.021; and that the new $\overline{R}$ is larger than the .6503 found for the two tests, 7 and 9. We include Test 3 in our battery, therefore, and proceed to calculate a_3, b_3 and c_3 (Table 66), following Step 11, in order to select a *fourth* test.

Step 14

The a_3 entries in Table 66 are the correlations of Test 3 with each of the other tests including the criterion. The criterion correlation is entered in the $-$ C column with a negative sign (i.e., as $-$ .530).

(1) The formula for b_3 is $b_3 = a_3 + b_1$ (given test) $\times c_1$ (*third* selected test) $+ b_2$ (given test) $\times c_2$ (*third* selected test). To illustrate,

For Test 1: $b_3 = .340 + (.300)(-.560) + (.217)(-.385)$
$= .088$

For Test 4: $b_3 = .630 + (.490)(-.560) + (.409)(-.385)$
$= .199$

Check the b_3 entries by Step 11 (5). (*a*) Note that the b_3 entry for the *third* selected test (Test 3) equals the Z_3 entry for Test 3 in Table 64, namely, .563. (*b*) The entry in the criterion column equals the V_3 entry of the *third* selected test (Test 3) in Table 63, i.e., $-$ .097. (*c*) The Check Sum entry (1.161) equals the sum of the entries in the b_3 row.

(2) The formula for c_3 is $b_3 \times$ the *negative reciprocal* of the b_3 entry for the *third* selected test (Test 3). The negative reciprocal of .563 is $-$ 1.776. To illustrate the calculation for Test 5, $c_3 = .146 \times -1.776 = -.259$. Check the c_3 entries by Step 11 (7). (*a*) The c_3 row entry of the *third* selected test (Test 3) equals $-$ 1.000. (*b*) The c_3 entry in the Check Sum column, namely, $-$ 2.062, equals the sum of the c_3 row. (*c*) The product of the b_3 and c_3

entries in the criterion column (namely, $-.097 \times .172$) equals the quotient $\left(\frac{V_3^2}{Z_3}\right)$ (i.e., .0167) in absolute value.

Step 15

Repeat Step 12 to find V_4 and Z_4. The formula for V_4 is $V_4 = V_3 + b_3$ (criterion) $\times c_3$ (each test). Also, the formula for Z_4 is $Z_3 + b_3$ (a given test) $\times c_3$ (same test). For Test 4, $V_4 = -.091 + (-.097)(-.353)$ or $-.057$; and $Z_4 = .559 + (.199)(-.353)$ or .489. The quotient, $\frac{V_4^2}{Z_4}$, equals $\frac{(-.057)^2}{.489}$ or .0066. While none of the V_4 entries is large, Test 4 has the largest $\frac{V_4^2}{Z_4}$ quotient, and hence is selected as our *fourth* test. Enter .0066 $\left(\frac{V_4^2}{Z_4}\right)$ in column b, row 4, of Table 65. Follow the procedure of Step 9 to get $\overline{R} = .6595$. Note that $\frac{(N-1)}{(N-m)}$ is 99/96 or 1.031; and that the new $\overline{R}$ is but slightly larger than the $\overline{R}$ of .6586 found for the three tests, 7, 9, and 3. When $\overline{R}$ decreases or fails to increase, there is no point in adding new tests to the battery. The increase in $\overline{R}$ is so small as a result of adding Test 4 that it is hardly profitable to enlarge our battery by a fifth test. We shall add a fifth test, however, in order to illustrate a further step in the selection process.

Step 16

To choose a *fifth* test, calculate a_4, b_4, and c_4, following Step 11, and enter the results in Table 66. The a_4 entries are the correlations of the *fourth* selected test (Test 4) with each of the other tests including the criterion (*with sign reversed*).

(1) The formula for b_4 may readily be written by analogy to the formulas for b_3 and b_2 as follows: $b_4 = a_4 + b_1$ (given test) $\times c_1$ (*fourth* selected test) $+ b_2$ (given test) $\times c_2$ (*fourth* selected test) $+ b_3$ (given test) $\times c_3$ (*fourth* selected test). To illustrate

For Test 6: $b_4 = .300 + (.400)(-\ .490) + (-\ .034)(-\ .492) + (.179)(-\ .353) = .058$

For Test 10: $b_4 = .560 + (.130)(-\ .490) + (.337)(-\ .492) + (.031)(-\ .353) = .324$

Check the b_4 entries by Step 11 (5). (*a*) The b_4 entry for the *fourth* selected test (Test 4) equals the Z_4 entry for Test 4 in Table 64, namely, .489. (*b*) The entry in the criterion column equals the V_4 entry of the *fourth* selected test (Test 4), i.e., $-\ .057$. (*c*) The Check Sum (.715) equals the sum of the entries in the b_4 row.

(2) To find the entries c_4, multiply each b_4 by the *negative reciprocal* of the b_4 entry for the *fourth* selected test (Test 4). The negative reciprocal of .489 is $-\ 2.045$. To illustrate,

For Test 1: $c_4 = -\ .145 \times -\ 2.045 = .297.$

Check the c_4 entries by Step 11 (7). (*a*) The c_4 row entry of the *fourth* selected test (Test 4) equals $-\ 1.000$. (*b*) The c_4 entry in the Check Sum column, namely, $-\ 1.462$, equals the sum of the c_4 row. (*c*) The product of the b_4 and c_4 entries in the criterion column (namely, $-\ .057 \times .117$) equals the quotient $\frac{V_4^2}{Z_4}$ (i.e., .0066) in absolute value.

Step 17

Repeat Step 12 to find V_5 and Z_5. $V_5 = V_4 + b_4$ (criterion) $\times\ c_4$ (each test); and $Z_5 = Z_4 + b_4$ (a given test) $\times\ c_4$ (same test). Test 8 has the largest $\left(\frac{V_5^2}{Z_5}\right)$ quotient (i.e., .0031) and this number is entered in column b, row 5 of Table 65. Following Step 9, we get $\overline{R} = .6573$. This multiple correlation coefficient is smaller than the preceding $\overline{R}$. We need go no further, therefore, as we have reached the point of diminishing returns and the addition of a sixth test will not increase the multiple $\overline{R}$. It may be noted that four (really three) tests constitute a bat-

tery which has the highest validity of any combination of tests chosen from our list of ten. The multiple $\overline{R}$ between the criterion and all ten tests would be somewhat lower — when corrected for chance error — than the $\overline{R}$ we have found for our battery of four tests. The Wherry-Doolittle method not only selects the most economical battery but saves a large amount of statistical work.

2. Calculation of the Multiple Regression Equation for Tests Selected by the Wherry-Doolittle Method

Steps involved in setting up a multiple regression equation for the tests selected in Table 63 may be set down as follows:

TABLE 67

	7	9	3	4	− C
C_1	− 1.000	− .410	− .560	− .490	.550
C_2		− 1.000	− .385	− .492	.389
C_3			− 1.000	− .353	.172
C_4				− 1.000	.117

Step 1

Draw up a work sheet like that shown in Table 67. Enter the C entries for the four selected tests (namely, 7, 9, 3, and 4) and for the criterion, following the order in which the tests were selected for the battery. When equated to zero, each row in Table 67 is an equation defining the *beta* weights.

For our four tests, the equations are

$$-1.000\beta_7 - .410\beta_9 - .560\beta_3 - .490\beta_4 + .550 = 0$$
$$-1.000\beta_9 - .385\beta_3 - .492\beta_4 + .389 = 0$$
$$-1.000\beta_3 - .353\beta_4 + .172 = 0$$
$$-1.000\beta_4 + .117 = 0$$

Step 2

Solve the fourth equation to find $\beta_4 = .117$.

Step 3

Substitute for $\beta_4 = .117$ in the third equation to get $\beta_3 = .131$.

Step 4

Substitute for β_3 and β_4 in the second equation to get $\beta_9 =$ .280. Finally, substitute for β_3, β_4, and β_9 in the first equation to get $\beta_7 = .305$.

Step 5

The regression equation for predicting the criterion from the four selected tests (7, 9, 3, and 4) may be written in standard score form by means of formula (95), page 421, as follows:

$$\bar{z}_c = \beta_7 z_7 + \beta_9 z_9 + \beta_3 z_3 + \beta_4 z_4$$

in which $\beta_7 = \beta_{c7.934}$; $\beta_9 = \beta_{c9.734}$; $\beta_3 = \beta_{c3.974}$; $\beta_4 = \beta_{c4.973}$. Substituting for the β's we have

$$\bar{z}_c = .305z_7 + .280z_9 + .131z_3 + .117z_4.$$

To predict the criterion score of any subject in our group, substitute his scores in Tests 7, 9, 3, and 4 (expressed as σ-scores) in this equation.

Step 6

To write the regression equation in score form the β's must be transformed into b's by means of formula (94), page 421, as follows:

$$b_7 = \frac{\sigma_c}{\sigma_7}\beta_7;\; b_9 = \frac{\sigma_c}{\sigma_9}\beta_9;\; b_3 = \frac{\sigma_c}{\sigma_3}\beta_3;\; b_4 = \frac{\sigma_c}{\sigma_4}\beta_4.$$

The σ's are the *SD*'s of the test scores: σ_7 of Test 7, σ_9 of Test 9, σ_c of the criterion, etc. In general, $b_p = \frac{\sigma_c}{\sigma_p}\beta_p$.

Step 7

The regression equation in score form may now be written

$$\bar{X}_c = b_7X_7 + b_9X_9 + b_3X_3 + b_4X_4 + K \qquad \text{*(90) p. 419}$$

and the $\sigma_{\text{est.}X_c} = \sigma_c\sqrt{1 - R^2_{c(7934)}}$ (58) p. 320

* This equation is not written for our four tests because means and *SD*'s are not given in Table 62.

3. Checking the β Weights and Multiple R

Step 1

The β weights may be checked by formula (99), page 425, in which R is expressed in terms of beta coefficients. In the present example, we have

$$R^2_{c(7934)} = \beta_7 r_{c7} + \beta_9 r_{c9} + \beta_3 r_{c3} + \beta_4 r_{c4}$$

in which c equals the criterion and the r's are the correlations between the criterion (c) and the Tests, 7, 9, 3, and 4. Substituting for the r's and β's (computed in the last section) we have

$$\begin{aligned} R^2_{c(7934)} &= .305 \times .550 + .280 \times .550 + .131 \times .530 + .117 \times .520 \\ &= .1678 + .1540 + .0694 + .0608 = .4520 \\ R_{c(7934)} &= .6723 \end{aligned}$$

From $R^2_{c(7934)}$ we know that our battery accounts for 45% of the variance of the criterion. Also (p. 426) our four tests (7, 9, 3, and 4) contribute 17%, 15%, 7%, and 6%, respectively, to the variance of the criterion.

Step 2

The R^2 of .4520 calculated above should equal $(1 - K^2)$ when K^2 is taken from column c, row 4 in Table 65. From Table 65 we find that $1 - K^2 = 1 - .5481$ or .4519 which checks the R^2 found above — and hence the β weights — very closely.

Step 3

It will be noted that the multiple correlation coefficient of .6723 found above is somewhat larger than the shrunken $\overline{R}$ of .6595 found between the criterion and our battery of four tests in Table 65. The multiple correlation coefficient obtained from a sample always tends — through the operation of chance errors — to be *larger* than the correlation in the population from which the sample was drawn, especially when N is small or the number of test variables large. For this reason, the calculated R must be "adjusted" in order to give us a better estimate of the corre-

lation in the population.* The relationship of the $\bar{R}$, corrected for chance errors, to the R as usually calculated, is given by the following equation:

$$\bar{R}^2 = \frac{(N-1)R^2 - (m-1)}{(N-m)} \qquad (100)\dagger$$

(*relation of* R *to* $\bar{\text{R}}$ *corrected for chance errors*)

Substituting .4520 for R^2, 99 for $(N - 1)$, 96 for $(N - m)$ and 3 for $(m - 1)$, we have from (100) that

$$\bar{R}^2 = \frac{99 \times .4520 - 3}{96} = .4349$$

and

$$\bar{R} = .6595 \text{ (see Table 65)}$$

The $\bar{R}$ of .6595 is the corrected multiple correlation between our criterion and test battery, or the multiple correlation coefficient estimated for the population from which our sample was drawn. In the present problem, shrinkage in multiple R is quite small (.6723 − .6595 = .0128) as the sample is fairly large and there are only four tests in the multiple regression equation.

II. Applications of Partial and Multiple Correlation

1. Partial Correlation in Analysis

Partial correlation may be of decided value as an aid in analyzing the part played by each of several factors in determining a total result. An illustration may be cited from the work of Cyril Burt.‡ Burt wished to find to what extent a child's M.A., as measured by the Binet test, influences his school attainment. His subjects were 300 children, seven to fourteen years old. For each child (1) an M.A. was determined, (2) his scholastic achievement as measured by edu-

* Ezekiel, M., *Methods of Correlation Analysis* (1941), pp. 323–324.
† Wherry, *op. cit.*, p. 451.
‡ Burt, Cyril, *Mental and Scholastic Tests* (London: 1921), pp. 180–184.

cational examinations and checked by teachers, and (3) his chronological age. The correlation between Binet M.A. and scholastic achievement (r_{12}) was .91. When chronological age (3) was partialled out the correlation ($r_{12.3}$) between Binet M.A. and scholastic achievement dropped to .68. This result shows, in the first place, that chronological age has a decided effect upon the correlation between M.A. and school work; it tends to increase or "dilate" the obtained r. This dilation is brought about by the fact that *both* M.A. and school attainment increase with C.A., and this common dependence on chronological age serves to boost the observed correlation. The residual partial correlation ($r_{12.3}$) of .68 indicates, however, a substantial relationship remaining between M.A. and school work when age is a constant factor. In other words, Binet M.A. is a substantial factor in a pupil's school attainment at each age level from seven to fourteen. Taking the analysis a step further, Burt found that the correlation between (2) school work and (3) chronological age (r_{23}) was .87; and that when Binet M.A. was held constant, the partial r ($r_{23.1}$) between school work and C.A. was reduced to .49. This persistence of a substantial relationship between school work and C.A., when variability arising from differences in M.A. is eliminated, offers confirmatory evidence according to Burt of the "undue influence of age upon school classification."

From analyses made through the elimination of factors by partial correlation, "causal" relationships may often be determined. Phillips,* for example, in a study of causes contributing to absence on account of illness among government employees over a period of a year, found that the observed correlation between absence and mean temperature on the day of absence was $-.37$. When the four factors (1) relative humidity at 8:00 A.M.; (2) relative humidity at noon of the previous day; (3) inches of rainfall on the day of absence; and (4) percent of possible sunshine on the day of absence were held constant, the partial correlation remaining between absence and temperature

* Phillips, F. E., Application of Partial Correlation to a Health Problem, *Public Health Report*, Reprint No. 867 (1923).

was − .39. Since the partial correlation between absence and temperature was the only r not reduced by the elimination of other factors the conclusion seems to be that of the factors studied, temperature on the day of absence is the most important contributing cause of absence. Illness, of course, must be taken as the primary cause of absence. It must be clearly understood that partial correlation has nothing to say about causal relations. One cannot say which of two variables is the cause and which the effect, when all one has is the correlation between them. Sometimes, however, cause and effect distinctions are a matter of common-sense analysis. In the illustration given above, for instance, the distinction between cause and effect is clear.

Another example of the use of partial correlation in a "causal" investigation is found in the work of Reavis.* This investigator undertook to ferret out the causes of attendance and non-attendance in rural schools. Certain factors (1) distance from school, (2) age-grade relationship, (3) kind of work done by pupils, (4) training and experience of the teacher, (5) school equipment, and (6) character of the community were selected as presumably having some effect upon school attendance. When partial correlation coefficients were calculated it was found that the original correlations between attendance and distance from school, and between attendance and character of the community, were the least reduced. The first coefficient was lowered from − .45 to − .43; and the second from .30 to .28. Of all of the factors selected, therefore, these two seemed to have the most direct and independent influence upon school attendance. As in the problem of temperature and absence, cited above, the distinction between cause and effect is clear: it is evident that distance from school and character of community are the causes and not the effects of good or poor school attendance.

* Reavis, George, Factors Controlling Attendance in Rural Schools, Teachers College, *Columbia University, Contributions to Education,* No. 108 (1920), 52–69.

2. Multiple Correlation in Analysis

Multiple correlation is often useful when one wishes to determine the influence of a number of test variables, taken singly and together, upon the criterion variable being studied. Also, as shown in Section I, multiple correlation enables us to select from a number of tests the most valid battery for forecasting a criterion of worker performance.* A few illustrations of the application of multiple correlation to psychological problems will be cited here; the student will encounter many in the literature. In a group of fifty-seven fourth-grade children,† the r between educational achievement and M.A. was .595. When physical efficiency (vigor, stamina, etc.) as estimated by teachers was added to M.A., the R of educational achievement with M.A. plus physical efficiency was .653, a gain of about .06 point. However, when emotional maturity (as estimated by teachers) was added to the battery M.A. plus physical efficiency, and still further social maturity (as estimated by teachers) was added to M.A. plus physical efficiency plus emotional maturity, the multiple correlation was unchanged. Gates concludes: "Physical fitness, then, appears to exert a greater specific influence (i.e., over and above the r with M.A.) upon achievement than does either social or emotional maturity or both combined. Both combined add practically nothing of value to a team of M.A. plus physical fitness for purposes of predicting scholastic achievement."

Burks‡ has made use of multiple correlation in determining the relative contribution of heredity and environment to a child's I.Q. as measured by Stanford-Binet. The R between I.Q. and parental intelligence test score *plus* environmental index (by Whittier Home Scale) was found to be .61 for an N of 105. Since

* Stead and Shartle, *op. cit.*, Chapters 5–9 inclusive.

† Gates, A. I., "The Nature and Educational Significance of Physical Status and of Mental, Physiological, Social and Emotional Maturity," *Journal of Educational Psychology*, 15 (1924), 347–349.

‡ Burks, B. S., The Relative Influence of Nature and Nurture upon Mental Development; a Comparative Study of Foster Parent-Foster Child Resemblance and True Parent-True Child Resemblance, *27th Yearbook, N.S.S.E.* (1928), Part I, 219–316.

R^2 is .37, about 37% of the variance of children's intelligence may be attributed to the combined effect of home environment and parents' mental level. Parental intelligence contributed 33%, and home environment 4%, to the 37% accounted for by these two factors. The remaining 63% is attributable to factors not measured by these two.

3. Limitations to the Use of Partial and Multiple Correlation

Certain limitations to the use of partial and multiple correlation may be indicated in concluding this section.

(1) In order that partial coefficients of correlation be valid measures of relationship, it is necessary that all zero order coefficients be computed from data in which the regression is *linear*. If there is any doubt as to linearity, the tests given on page 372 should be employed.

(2) The number of cases in a multiple correlation problem should be large, especially if there are a number of variables; otherwise the coefficients calculated from the data will have little significance. Coefficients which are misleadingly high or low may be obtained when studies which involve many variables are based on relatively few cases. The question of accuracy of computation is also involved. A general rule advocated by many workers is that results should be carried to as many decimals as there are variables in the problem. How strictly this rule is to be followed must depend upon the accuracy of the original measures.

(3) A serious limitation to a clear-cut interpretation of a partial r arises from the fact that most of the tests employed by psychologists probably depend upon a large number of "determiners." When we "partial out" the influence of clear-cut and relatively objective factors such as age, height, school grade, etc., we have a reasonably clear notion of what the "partials" mean. But when we attempt to render variability due to "logical memory" constant by partialling out memory test scores from the correlation between general intelligence test scores and educational achievement, the result is by no

means so unequivocal. The abilities determining the scores in general intelligence *and* in school achievement undoubtedly overlap the memory test in other respects than in the "memory" involved. Partialling out a memory test score from the correlation between general intelligence and educational achievement, therefore, will render constant the influence of many factors not strictly "memory," i.e., partial out too much.*

To illustrate this point again it would be fallacious to interpret the partial correlation between reading comprehension and arithmetic, say, with the influence of "general intelligence" partialled out, as giving the net relationship between these two variables for a "constant" degree of intelligence. Both reading and arithmetic enter with heavy, but unknown, weight into most general intelligence tests; hence the partial correlation between these two, for general intelligence constant, cannot be interpreted in a clear-cut and meaningful way.

Partial r's obtained from psychological and educational tests, though often difficult to interpret, may be used in multiple regression equations when the purpose is to determine the relative weight to be assigned the various tests of a battery. But we should be cautious in attempting to give psychological meaning to such residual, i.e., partial, r's. Several writers have discussed this problem, and should be referred to by the investigator who plans to use partial and multiple correlation extensively.†

(4) Perhaps the chief limitation to R, the coefficient of multiple correlation, is the fact that, since it is always positive, variable errors of sampling tend to accumulate and thus make the coefficient too large. A correction to be applied to R, when the sample is small and the number of variables large, has been given on page 451. This correction gives the value which R would most probably take in the population from which our sample was drawn.

* Burks, B. S., Statistical Hazards in Nature-Nurture Investigations, *27th Yearbook, N.S.S.E.* (1928), Part I, 9–33.

† Burks, B. S., "On the Inadequacy of the Partial and Multiple Correlation Technique," *Journal of Educational Psychology*, 17 (1926), 532–540.

Moore, T. V., Partial Correlations, *Studies in Psychology and Psychiatry from the Catholic University of America*, 3 (1932), 1–39.

PROBLEMS

1. The following data* were assembled for sixteen large cities (of around 500,000 inhabitants) in a study of factors making for variation in crime.

X_c (criterion) = crime rate: number known offenses per 1000 inhabitants
X_1 = percentage of male inhabitants
X_2 = percentage of male native whites of native parentage
X_3 = percentage of foreign-born males
X_4 = number children under five per 1000 married women fifteen to forty-four years old
X_5 = number Negroes per 100 of population
X_6 = number male children of foreign-born parents per 100 of population
X_7 = number males and females ten years and over in manufacturing per 100 of population

$M_c = 19.9$ $M_1 = 49.2$ $M_2 = 22.8$ $M_3 = 10.2$ $M_4 = 481.4$ $M_5 = 4.7$
$\sigma_c = 7.9$ $\sigma_1 = 1.3$ $\sigma_2 = 7.2$ $\sigma_3 = 4.6$ $\sigma_4 = 74.4$ $\sigma_5 = 4.0$
$M_6 = 13.1$ $M_7 = 21.7$
$\sigma_6 = 4.2$ $\sigma_7 = 4.3$

Intercorrelations

	1	2	3	4	5	6	7
C	.44	.44	− .34	− .31	.51	− .54	− .20
1		.01	.25	− .19	− .15	.01	.22
2			− .92	− .54	.55	− .93	− .30
3				.44	− .68	.82	.40
4					− .06	.52	.74
5						− .67	− .14
6							.21

(*a*) By means of the Wherry-Doolittle method select those variables which give a maximum correlation with the criterion.

(*b*) Work out the regression equation in score form (p. 419) and $\sigma_{(\text{est. } X_C)}$.

* Ogburn, W. F., "Factors in the Variation of Crime among Cities," *Journal of the American Statistical Association*, 30 (1935), 12–34.

(*c*) Determine the independent contribution of each of the selected factors to crime rate (to R^2).

(*d*) Compare R and $\bar{R}$. Why is the adjustment fairly large? (see p. 451)

2. (*a*) What is the probable crime rate (from Problem 1) for a city in which $X_6 = 15.0$, $X_1 = 50\%$, $X_5 = 6.0$ and $X_7 = 20.00$?

(*b*) For a city in which $X_6 = 13$, $X_1 = 48\%$, $X_5 = 5.0$ and $X_7 = 22.00$?

(*c*) By how much does the use of multiple R reduce $\sigma_{(\text{est. } X_c)}$?

3. In Problem 4, page 432:

(*a*) Work out the regression equation using the Wherry-Doolittle method.

(*b*) How much shrinkage is there when $R_{1(23)}$ is corrected for chance errors (p. 451)?

Answers

1. (*a*) The $\bar{R}$'s are, for Test 6, .540; for Tests 6 and 1, .674; for Tests 6, 1, and 5, .713; for Tests 6, 1, 5, and 7, .722. $\bar{R}$ drops to .702, when Test 4 is added.

(*b*) $\bar{X}_c = -.42\, X_6 + 3.35\, X_1 + .82\, X_5 - .40\, X_7 - 134.59.$

$\sigma_{(\text{est. } X_c)} = 5.47$

(*c*) $R^2_{c(6157)} = .121 + .242 + .210 + .043$. Tests 6, 1, 5, and 7 contribute 12%, 24%, 21%, and 4%, respectively.

(*d*) $R = .785$; $\bar{R} = .722$; shrinkage is .063.

2. (*a*) 23.53

(*b*) 16.05

(*c*) From 7.9 to 5.5 or 30%

3. (*b*) $\bar{R}_{1(23)}$ is .59.

REFERENCE TABLES

TABLE 17

Fractional Parts of the Total Area (Taken as 10,000) under the Normal Probability Curve, Corresponding to Distances on the Baseline between the Mean and Successive Points Laid Off from the Mean in Units of Standard Deviation

Example: between the mean and a point 1.38σ $\left(\frac{x}{\sigma} = 1.38\right)$ are found 41.62% of the entire area under the curve.

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0.7	2580	2611	2642	2673	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3290	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4383	4394	4406	4418	4429	4441
1.6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2.0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2.9	4981	4982	4982	4983	4984	4984	4985	4985	4986	4986
3.0	4986.5	4986.9	4987.4	4987.8	4988.2	4988.6	4988.9	4989.3	4989.7	4990.0
3.1	4990.3	4990.6	4991.0	4991.3	4991.6	4991.8	4992.1	4992.4	4992.6	4992.9
3.2	4993.129									
3.3	4995.166									
3.4	4996.631									
3.5	4997.674									
3.6	4998.409									
3.7	4998.922									
3.8	4999.277									
3.9	4999.519									
4.0	4999.683									
4.5	4999.966									
5.0	4999.997133									

TABLE 18

Fractional Parts of the Total Area (Taken as 10,000) under the Normal Probability Curve, Corresponding to Distances on the Baseline between the Mean and Successive Points Laid Off from the Mean in Units of *PE*

Example: between the mean and a point 1.55 PE $\left(\frac{x}{PE} = 1.55\right)$ from the mean are found 35.21% of the entire area under the curve.

$\frac{x}{PE}$	.00	.05	$\frac{x}{PE}$	.00	.05
0	0000	0135	3.0	4785	4802
.1	0269	0403	3.1	4817	4832
.2	0537	0670	3.2	4846	4858
.3	0802	0933	3.3	4870	4881
.4	1063	1193	3.4	4891	4900
.5	1320	1447	3.5	4909	4917
.6	1571	1695	3.6	4924	4931
.7	1816	1935	3.7	4937	4943
.8	2053	2168	3.8	4948	4953
.9	2281	2392	3.9	4957	4961
1.0	2500	2606	4.0	4965	4968
1.1	2709	2810	4.1	4972	4974
1.2	2909	3004	4.2	4977	4979
1.3	3097	3187	4.3	4981	4983
1.4	3275	3360	4.4	4985	4987
1.5	3442	3521	4.5	4988	4989
1.6	3597	3671	4.6	4990	4991
1.7	3742	3811	4.7	4992	4993
1.8	3876	3939	4.8	4994	4995
1.9	4000	4058	4.9	4995	4996
2.0	4113	4166	5.0	4996	4997
2.1	4217	4265	5.1	4997.1	4997.4
2.2	4311	4354	5.2	4997.7	4998
2.3	4396	4435	5.3	4998.2	4998.5
2.4	4473	4508	5.4	4998.6	4998.8
2.5	4541	4573	5.5	4999	4999.1
2.6	4603	4631	5.6	4999.2	4999.3
2.7	4657	4682	5.7	4999.4	4999.5
2.8	4705	4727	5.8	4999.54	4999.6
2.9	4748	4767	5.9	4999.65	4999.7

TABLE 23

To Facilitate the Calculation of *T*-Scores

The percents refer to the percentage of the total frequency below a given score + 1/2 of the frequency on that score. *T*-scores are read directly from the given percentages.

Percent	*T-score*	*Percent*	*T-score*
.0032	10	53.98	51
.0048	11	57.93	52
.007	12	61.79	53
.011	13	65.54	54
.016	14	69.15	55
.023	15	72.57	56
.034	16	75.80	57
.048	17	78.81	58
.069	18	81.59	59
.097	19	84.13	60
.13	20	86.43	61
.19	21	88.49	62
.26	22	90.32	63
.35	23	91.92	64
.47	24	93.32	65
.62	25	94.52	66
.82	26	95.54	67
1.07	27	96.41	68
1.39	28	97.13	69
1.79	29	97.72	70
2.28	30	98.21	71
2.87	31	98.61	72
3.59	32	98.93	73
4.46	33	99.18	74
5.48	34	99.38	75
6.68	35	99.53	76
8.08	36	99.65	77
9.68	37	99.74	78
11.51	38	99.81	79
13.57	39	99.865	80
15.87	40	99.903	81
18.41	41	99.931	82
21.19	42	99.952	83
24.20	43	99.966	84
27.43	44	99.977	85
30.85	45	99.984	86
34.46	46	99.9890	87
38.21	47	99.9928	88
42.07	48	99.9952	89
46.02	49	99.9968	90
50.00	50		

TABLE 29

Table of t

For Use in Determining the Reliability of Statistics.
If N is Large, Tables 17 and 18 May Be Used.

Example: An $(N - 1) = 35$ and $t = 2.03$ means that 5 times in 100 trials a divergence as large as that obtained may be expected *in the positive and negative directions.*

Degrees of Freedom $(N - 1)$	PROBABILITY (P) 0.50	0.10	0.05	0.02	0.01
1	$t =$ 1.000	$t =$ 6.34	$t =$ 12.71	$t =$ 31.82	$t =$ 63.66
2	0.816	2.92	4.30	6.96	9.92
3	.765	2.35	3.18	4.54	5.84
4	.741	2.13	2.78	3.75	4.60
5	.727	2.02	2.57	3.36	4.03
6	.718	1.94	2.45	3.14	3.71
7	.711	1.90	2.36	3.00	3.50
8	.706	1.86	2.31	2.90	3.36
9	.703	1.83	2.26	2.82	3.25
10	.700	1.81	2.23	2.76	3.17
11	.697	1.80	2.20	2.72	3.11
12	.695	1.78	2.18	2.68	3.06
13	.694	1.77	2.16	2.65	3.01
14	.692	1.76	2.14	2.62	2.98
15	.691	1.75	2.13	2.60	2.95
16	.690	1.75	2.12	2.58	2.92
17	.689	1.74	2.11	2.57	2.90
18	.688	1.73	2.10	2.55	2.88
19	.688	1.73	2.09	2.54	2.86
20	.687	1.72	2.09	2.53	2.84
21	.686	1.72	2.08	2.52	2.83
22	.686	1.72	2.07	2.51	2.82
23	.685	1.71	2.07	2.50	2.81
24	.685	1.71	2.06	2.49	2.80
25	.684	1.71	2.06	2.48	2.79
26	.684	1.71	2.06	2.48	2.78
27	.684	1.70	2.05	2.47	2.77
28	.683	1.70	2.05	2.47	2.76
29	.683	1.70	2.04	2.46	2.76
30	.683	1.70	2.04	2.46	2.75
35	.682	1.69	2.03	2.44	2.72
40	.681	1.68	2.02	2.42	2.71
45	.680	1.68	2.02	2.41	2.69
50	.679	1.68	2.01	2.40	2.68
60	.678	1.67	2.00	2.39	2.66
70	.678	1.67	2.00	2.38	2.65
80	.677	1.66	1.99	2.38	2.64
90	.677	1.66	1.99	2.37	2.63

TABLE 32

Table of Chi-Square

(The values of χ^2 are printed in the body of the table.)

Adapted from R. A. Fisher's *Statistical Method for Research Workers*, Oliver & Boyd, by permission of publishers.

df	P = 0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01
1	0.000157	0.000628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892

TABLE 49

Correlation Coefficients at the 5% and 1% Levels of Significance

Example: When N is 52 and $(N - 2)$ is 50, an r must be .273 to be significant at .05 level, and .354 to be significant at .01 level.

Degrees of freedom $(N - 2)$	.05	.01	Degrees of freedom $(N - 2)$	.05	.01
1	.997	1.000	24	.388	.496
2	.950	.990	25	.381	.487
3	.878	.959	26	.374	.478
4	.811	.917	27	.367	.470
5	.754	.874	28	.361	.463
6	.707	.834	29	.355	.456
7	.666	.798	30	.349	.449
8	.632	.765	35	.325	.418
9	.602	.735	40	.304	.393
10	.576	.708	45	.288	.372
11	.553	.684	50	.273	.354
12	.532	.661	60	.250	.325
13	.514	.641	70	.232	.302
14	.497	.623	80	.217	.283
15	.482	.606	90	.205	.267
16	.468	.590	100	.195	.254
17	.456	.575	125	.174	.228
18	.444	.561	150	.159	.208
19	.433	.549	200	.138	.181
20	.423	.537	300	.113	.148
21	.413	.526	400	.098	.128
22	.404	.515	500	.088	.115
23	.396	.505	1000	.062	.081

TABLE 54

DEVIATES (x/σ) IN TERMS OF σ-UNITS AND ORDINATES (z) FOR GIVEN AREAS MEASURED FROM THE MEAN OF A NORMAL DISTRIBUTION WHOSE TOTAL AREA = 1.00

[$x/\sigma = x$]

Area from the Mean (α)	x or (x/σ)	z	Area from the Mean (α)	x or (x/σ)	z
.00	.000	.399	.26	.706	311
.01	.025	.399	27	.739	.304
.02	.050	.398	.28	.772	.296
.03	.075	.398	.29	.806	.288
.04	.100	.397	.30	.842	.280
.05	.126	.396	.31	.878	.271
.06	.151	.394	.32	.915	.262
.07	.176	.393	.33	.954	.253
.08	.202	.391	.34	.995	.243
09	.228	.389	.35	1.036	.233
.10	.253	.386	.36	1.080	.223
.11	.279	.384	.37	1.126	.212
.12	.305	.381	.38	1.175	.200
.13	.332	.378	.39	1.227	.188
.14	.358	.374	.40	1.282	.176
.15	.385	.370	.41	1.341	.162
16	.412	.366	.42	1.405	.149
.17	.440	.362	.43	1.476	.134
.18	.468	.358	.44	1.555	.119
.19	.496	.353	.45	1.645	.103
.20	.524	.348	.46	1.751	.086
.21	.553	.342	.47	1.881	.068
.22	.583	.337	.48	2.054	.048
.23	.613	.331	.49	2.326	.027
.24	.643	.324	.50	∞	.000
.25	.675	.318			

* At the .05 level tne CR = 1.98, at the .01 level 2.63, when the (N − 1) = 99

TABLE 60

A Table to Infer the Value of $\sqrt{1 - r^2}$ from a Given Value of r

r	$\sqrt{1 - r^2}$	r	$\sqrt{1 - r^2}$	r	$\sqrt{1 - r^2}$
.0000	1.0000	.3400	.9404	.6800	.7332
.01	9999	.35	.9367	.69	.7238
.02	.9998	.36	.9330	.70	.7141
.03	9995	.37	.9290	.71	.7042
.04	.9992	.38	.9250	.72	.6940
.05	.9987	.39	.9208	.73	.6834
.06	.9982	.40	.9165	.74	.6726
.07	.9975	.41	.9121	.75	.6614
.08	.9968	.42	.9075	.76	.6499
.09	.9959	.43	.9028	.77	.6380
.10	.9950	.44	.8980	.78	.6258
.11	.9939	.45	.8930	.79	.6131
.12	.9928	.46	.8879	.80	.6000
.13	.9915	.47	.8827	.81	.5864
.14	.9902	.48	.8773	.82	.5724
.15	.9887	.49	.8717	.83	.5578
.16	.9871	.50	.8660	.84	.5426
17	.9854	.51	.8617	.85	.5268
.18	.9837	.52	.8542	.86	.5103
.19	.9818	.53	.8480	.87	.4931
.20	.9798	.54	.8417	.88	.4750
.21	.9777	.55	.8352	.89	.4560
.22	.9755	.56	.8285	.90	.4359
.23	.9732	.57	.8216	.91	.4146
.24	.9708	.58	.8146	.92	.3919
.25	.9682	.59	.8074	.93	.3676
.26	.9650	.60	.8000	.94	.3412
.27	.9629	.61	.7924	.95	.3122
.28	.9600	.62	.7846	.96	.2800
.29	.9570	.63	.7766	.97	.2431
.30	.9539	.64	.7684	.98	.1990
.31	.9507	.65	.7599	.99	.1411
.32	.9474	.66	.7513	1.00	.0000
.33	.9440	.67	.7424		

TABLE OF SQUARES AND SQUARE ROOTS OF THE NUMBERS FROM 1 TO 1000

TABLE OF SQUARES AND SQUARE ROOTS OF THE NUMBERS FROM 1 TO 1000

Number	Square	Square Root	Number	Square	Square Root
1	1	1.000	51	26 01	7.141
2	4	1.414	52	27 04	7.211
3	9	1.732	53	28 09	7.280
4	16	2.000	54	29 16	7.348
5	25	2.236	55	30 25	7.416
6	36	2.449	56	31 36	7.483
7	49	2.646	57	32 49	7.550
8	64	2.828	58	33 64	7.616
9	81	3.000	59	34 81	7.681
10	1 00	3.162	60	36 00	7.746
11	1 21	3.317	61	37 21	7.810
12	1 44	3.464	62	38 44	7.874
13	1 69	3.606	63	39 69	7.937
14	1 96	3.742	64	40 96	8.000
15	2 25	3.873	65	42 25	8 062
16	2 56	4.000	66	43 56	8.124
17	2 89	4.123	67	44 89	8.185
18	3 24	4.243	68	46 24	8.246
19	3 61	4.359	69	47 61	8.307
20	4 00	4.472	70	49 00	8.367
21	4 41	4.583	71	50 41	8.426
22	4 84	4.690	72	51 84	8.485
23	5 29	4.796	73	53 29	8.544
24	5 76	4.899	74	54 76	8.602
25	6 25	5.000	75	56 25	8.660
26	6 76	5.099	76	57 76	8.718
27	7 29	5.196	77	59 29	8.775
28	7 84	5.292	78	60 84	8.832
29	8 41	5.385	79	62 41	8.888
30	9 00	5.477	80	64 00	8.944
31	9 61	5.568	81	65 61	9.000
32	10 24	5.657	82	67 24	9.055
33	10 89	5.745	83	68 89	9.110
34	11 56	5.831	84	70 56	9.165
35	12 25	5.916	85	72 25	9.220
36	12 96	6.000	86	73 96	9.274
37	13 69	6.083	87	75 69	9.327
38	14 44	6.164	88	77 44	9.381
39	15 21	6.245	89	79 21	9.434
40	16 00	6.325	90	81 00	9.487
41	16 81	6.403	91	82 81	9.539
42	17 64	6.481	92	84 64	9.592
43	18 49	6.557	93	86 49	9.644
44	19 36	6.633	94	88 36	9.695
45	20 25	6.708	95	90 25	9.747
46	21 16	6.782	96	92 16	9.798
47	22 09	6.856	97	94 09	9.849
48	23 04	6.928	98	96 04	9 899
49	24 01	7.000	99	98 01	9 950
50	25 00	7.071	100	1 00 00	10 000

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
101	1 02 01	10.050	151	2 28 01	12.288
102	1 04 04	10.100	152	2 31 04	12.329
103	1 06 09	10.149	153	2 34 09	12.369
104	1 08 16	10.198	154	2 37 16	12.410
105	1 10 25	10.247	155	2 40 25	12.450
106	1 12 36	10.296	156	2 43 36	12.490
107	1 14 49	10.344	157	2 46 49	12.530
108	1 16 64	10.392	158	2 49 64	12.570
109	1 18 81	10.440	159	2 52 81	12.610
110	1 21 00	10.488	160	2 56 00	12.649
111	1 23 21	10.536	161	2 59 21	12.689
112	1 25 44	10.583	162	2 62 44	12.728
113	1 27 69	10.630	163	2 65 69	12.767
114	1 29 96	10.677	164	2 68 96	12.806
115	1 32 25	10.724	165	2 72 25	12.845
116	1 34 56	10.770	166	2 75 56	12.884
117	1 36 89	10.817	167	2 78 89	12.923
118	1 39 24	10.863	168	2 82 24	12.961
119	1 41 61	10.909	169	2 85 61	13.000
120	1 44 00	10.954	170	2 89 00	13.038
121	1 46 41	11.000	171	2 92 41	13.077
122	1 48 84	11.045	172	2 95 84	13.115
123	1 51 29	11.091	173	2 99 29	13.153
124	1 53 76	11.136	174	3 02 76	13.191
125	1 56 25	11.180	175	3 06 25	13.229
126	1 58 76	11.225	176	3 09 76	13.266
127	1 61 29	11.269	177	3 13 29	13.304
128	1 63 84	11.314	178	3 16 84	13.342
129	1 66 41	11.358	179	3 20 41	13.379
130	1 69 00	11.402	180	3 24 00	13.416
131	1 71 61	11.446	181	3 27 61	13.454
132	1 74 24	11.489	182	3 31 24	13.491
133	1 76 89	11.533	183	3 34 89	13.528
134	1 79 56	11.576	184	3 38 56	13.565
135	1 82 25	11.619	185	3 42 25	13.601
136	1 84 96	11.662	186	3 45 96	13.638
137	1 87 69	11.705	187	3 49 69	13.675
138	1 90 44	11.747	188	3 53 44	13.711
139	1 93 21	11.790	189	3 57 21	13.748
140	1 96 00	11.832	190	3 61 00	13.784
141	1 98 81	11.874	191	3 64 81	13.820
142	2 01 64	11.916	192	3 68 64	13.856
143	2 04 49	11.958	193	3 72 49	13.892
144	2 07 36	12.000	194	3 76 36	13.928
145	2 10 25	12.042	195	3 80 25	13.964
146	2 13 16	12.083	196	3 84 16	14.000
147	2 16 09	12.124	197	3 88 09	14.036
148	2 19 04	12.166	198	3 92 04	14.071
149	2 22 01	12.207	199	3 96 01	14.107
150	2 25 00	12.247	200	4 00 00	14.142

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
201	4 04 01	14.177	251	6 30 01	15.843
202	4 08 04	14.213	252	6 35 04	15.875
203	4 12 09	14.248	253	6 40 09	15.906
204	4 16 16	14.283	254	6 45 16	15.937
205	4 20 25	14.318	255	6 50 25	15.969
206	4 24 36	14.353	256	6 55 36	16.000
207	4 28 49	14.387	257	6 60 49	16.031
208	4 32 64	14.422	258	6 65 64	16.062
209	4 36 81	14.457	259	6 70 81	16.093
210	4 41 00	14.491	260	6 76 00	16.125
211	4 45 21	14.526	261	6 81 21	16.155
212	4 49 44	14.560	262	6 86 44	16.186
213	4 53 69	14.595	263	6 91 69	16.217
214	4 57 96	14.629	264	6 96 96	16.248
215	4 62 25	14.663	265	7 02 25	16.279
216	4 66 56	14.697	266	7 07 56	16.310
217	4 70 89	14.731	267	7 12 89	16.340
218	4 75 24	14.765	268	7 18 24	16.371
219	4 79 61	14.799	269	7 23 61	16.401
220	4 84 00	14.832	270	7 29 00	16.432
221	4 88 41	14.866	271	7 34 41	16.462
222	4 92 84	14.900	272	7 39 84	16.492
223	4 97 29	14.933	273	7 45 29	16.523
224	5 01 76	14.967	274	7 50 76	16.553
225	5 06 25	15.000	275	7 56 25	16.583
226	5 10 76	15.033	276	7 61 76	16.613
227	5 15 29	15.067	277	7 67 29	16.643
228	5 19 84	15.100	278	7 72 84	16.673
229	5 24 41	15.133	279	7 78 41	16.703
230	5 29 00	15.166	280	7 84 00	16.733
231	5 33 61	15.199	281	7 89 61	16.763
232	5 38 24	15.232	282	7 95 24	16.793
233	5 42 89	15.264	283	8 00 89	16.823
234	5 47 56	15.297	284	8 06 56	16.852
235	5 52 25	15.330	285	8 12 25	16.882
236	5 56 96	15.362	286	8 17 96	16.912
237	5 61 69	15.395	287	8 23 69	16.941
238	5 66 44	15.427	288	8 29 44	16.971
239	5 71 21	15.460	289	8 35 21	17.000
240	5 76 00	15.492	290	8 41 00	17.029
241	5 80 81	15.524	291	8 46 81	17.059
242	5 85 64	15.556	292	8 52 64	17.088
243	5 90 49	15.588	293	8 58 49	17.117
244	5 95 36	15.620	294	8 64 36	17.146
245	6 00 25	15.652	295	8 70 25	17.176
246	6 05 16	15.684	296	8 76 16	17.205
247	6 10 09	15.716	297	8 82 09	17.234
248	6 15 04	15.748	298	8 88 04	17.263
249	6 20 01	15.780	299	8 94 01	17.292
250	6 25 00	15.811	300	9 00 00	17.321

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
301	9 06 01	17.349	351	12 32 01	18.735
302	9 12 04	17.378	352	12 39 04	18.762
303	9 18 09	17.407	353	12 46 09	18.788
304	9 24 16	17.436	354	12 53 16	18.815
305	9 30 25	17.464	355	12 60 25	18.841
306	9 36 36	17.493	356	12 67 36	18.868
307	9 42 49	17.521	357	12 74 49	18 894
308	9 48 64	17.550	358	12 81 64	18.921
309	9 54 81	17.578	359	12 88 81	18.947
310	9 61 00	17.607	360	12 96 00	18.974
311	9 67 21	17 635	361	13 03 21	19.000
312	9 73 44	17 664	362	13 10 44	19.026
313	9 79 69	17.692	363	13 17 69	19.053
314	9 85 96	17.720	364	13 24 96	19.079
315	9 92 25	17.748	365	13 32 25	19.105
316	9 98 56	17.776	366	13 39 56	19.131
317	10 04 89	17.804	367	13 46 89	19.157
318	10 11 24	17 833	368	13 54 24	19.183
319	10 17 61	17.861	369	13 61 61	19.209
320	10 24 00	17.889	370	13 69 00	19.235
321	10 30 41	17.916	371	13 76 41	19.261
322	10 36 84	17 944	372	13 83 84	19.287
323	10 43 29	17 972	373	13 91 29	19.313
324	10 49 76	18.000	374	13 98 76	19.339
325	10 56 25	18.028	375	14 06 25	19.363
326	10 62 76	18.055	376	14 13 76	19.391
327	10 69 29	18.083	377	14 21 29	19.416
328	10 75 84	18.111	378	14 28 84	19.442
329	10 82 41	18.138	379	14 36 41	19.468
330	10 89 00	18.166	380	14 44 00	19.494
331	10 95 61	18.193	381	14 51 61	19.519
332	11 02 24	18.221	382	14 59 24	19.545
333	11 08 89	18.248	383	14 66 89	19.570
334	11 15 56	18.276	384	14 74 56	19.596
335	11 22 25	18.303	385	14 82 25	19.621
336	11 28 96	18.330	386	14 89 96	19.647
337	11 35 69	18.358	387	14 97 69	19.672
338	11 42 44	18.385	388	15 05 44	19.698
339	11 49 21	18.412	389	15 13 21	19.723
340	11 56 00	18.439	390	15 21 00	19.748
341	11 62 81	18.466	391	15 28 81	19.774
342	11 69 64	18.493	392	15 36 64	19.799
343	11 76 49	18.520	393	15 44 49	19.824
344	11 83 36	18.547	394	15 52 36	19.849
345	11 90 25	18.574	395	15 60 25	19.875
346	11 97 16	18.601	396	15 68 16	19.900
347	12 04 09	18.628	397	15 76 09	19.925
348	12 11 04	18.655	398	15 84 04	19.950
349	12 18 01	18.682	399	15 92 01	19.975
350	12 25 00	18.708	400	16 00 00	20.000

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
401	16 08 01	20.025	451	20 34 01	21.237
402	16 16 04	20.050	452	20 43 04	21.260
403	16 24 09	20.075	453	20 52 09	21.284
404	16 32 16	20.100	454	20 61 16	21.307
405	16 40 25	20.125	455	20 70 25	21.331
406	16 48 36	20.149	456	20 79 36	21.354
407	16 56 49	20.174	457	20 88 49	21.378
408	16 64 64	20.199	458	20 97 64	21.401
409	16 72 81	20.224	459	21 06 81	21.424
410	16 81 00	20.248	460	21 16 00	21.448
411	16 89 21	20.273	461	21 25 21	21.471
412	16 97 44	20.298	462	21 34 44	21.494
413	17 05 69	20.322	463	21 43 69	21.517
414	17 13 96	20.347	464	21 52 96	21.541
415	17 22 25	20.372	465	21 62 25	21.564
416	17 30 56	20.396	466	21 71 56	21.587
417	17 38 89	20.421	467	21 80 89	21.610
418	17 47 24	20.445	468	21 90 24	21.633
419	17 55 61	20.469	469	21 99 61	21.656
420	17 64 00	20.494	470	22 09 00	21.679
421	17 72 41	20.518	471	22 18 41	21.703
422	17 80 84	20.543	472	22 27 84	21.726
423	17 89 29	20.567	473	22 37 29	21.749
424	17 97 76	20.591	474	22 46 76	21.772
425	18 06 25	20.616	475	22 56 25	21.794
426	18 14 76	20.640	476	22 65 76	21.817
427	18 23 29	20.664	477	22 75 29	21.840
428	18 31 84	20.688	478	22 84 84	21.863
429	18 40 41	20.712	479	22 94 41	21.886
430	18 49 00	20.736	480	23 04 00	21.909
431	18 57 61	20.761	481	23 13 61	21.932
432	18 66 24	20.785	482	23 23 24	21.954
433	18 74 89	20.809	483	23 32 89	21.977
434	18 83 56	20.833	484	23 42 56	22.000
435	18 92 25	20.857	485	23 52 25	22.023
436	19 00 96	20.881	486	23 61 96	22.045
437	19 09 69	20.905	487	23 71 69	22.068
438	19 18 44	20.928	488	23 81 44	22.091
439	19 27 21	20.952	489	23 91 21	22.113
440	19 36 00	20.976	490	24 01 00	22.136
441	19 44 81	21.000	491	24 10 81	22.159
442	19 53 64	21.024	492	24 20 64	22.181
443	19 62 49	21.048	493	24 30 49	22.204
444	19 71 36	21.071	494	24 40 36	22.226
445	19 80 25	21.095	495	24 50 25	22.249
446	19 89 16	21.119	496	24 60 16	22.271
447	19 98 09	21.142	497	24 70 09	22.293
448	20 07 04	21.166	498	24 80 04	22.316
449	20 16 01	21.190	499	24 90 01	22.338
450	20 25 00	21.213	500	25 00 00	22.361

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
501	25 10 01	22.383	551	30 36 01	23.473
502	25 20 04	22.405	552	30 47 04	23.495
503	25 30 09	22.428	553	30 58 09	23.516
504	25 40 16	22.450	554	30 69 16	23.537
505	25 50 25	22.472	555	30 80 25	23.558
506	25 60 36	22.494	556	30 91 36	23.580
507	25 70 49	22.517	557	31 02 49	23.601
508	25 80 64	22.539	558	31 13 64	23.622
509	25 90 81	22.561	559	31 24 81	23.643
510	26 01 00	22.583	560	31 36 00	23.664
511	26 11 21	22.605	561	31 47 21	23.685
512	26 21 44	22.627	562	31 58 44	23.707
513	26 31 69	22.650	563	31 69 69	23.728
514	26 41 96	22.672	564	31 80 96	23.749
515	26 52 25	22.694	565	31 92 25	23.770
516	26 62 56	22.716	566	32 03 56	23.791
517	26 72 89	22.738	567	32 14 89	23.812
518	26 83 24	22.760	568	32 26 24	23.833
519	26 93 61	22.782	569	32 37 61	23.854
520	27 04 00	22.804	570	32 49 00	23.875
521	27 14 41	22.825	571	32 60 41	23.896
522	27 24 84	22.847	572	32 71 84	23.917
523	27 35 29	22.869	573	32 83 29	23.937
524	27 45 76	22.891	574	32 94 76	23.958
525	27 56 25	22.913	575	33 06 25	23.979
526	27 66 76	22.935	576	33 17 76	24.000
527	27 77 29	22.956	577	33 29 29	24.021
528	27 87 84	22.978	578	33 40 84	24.042
529	27 98 41	23.000	579	33 52 41	24.062
530	28 09 00	23.022	580	33 64 00	24.083
531	28 19 61	23.043	581	33 75 61	24.104
532	28 30 24	23.065	582	33 87 24	24.125
533	28 40 89	23.087	583	33 98 89	24.145
534	28 51 56	23.108	584	34 10 56	24.166
535	28 62 25	23.130	585	34 22 25	24.187
536	28 72 96	23.152	586	34 33 96	24.207
537	28 83 69	23.173	587	34 45 69	24.228
538	28 94 44	23.195	588	34 57 44	24.249
539	29 05 21	23.216	589	34 69 21	24.269
540	29 16 00	23.238	590	34 81 00	24.290
541	29 26 81	23.259	591	34 92 81	24.310
542	29 37 64	23.281	592	35 04 64	24.331
543	29 48 49	23.302	593	35 16 49	24.352
544	29 59 36	23.324	594	35 28 36	24.372
545	29 70 25	23.345	595	35 40 25	24.393
546	29 81 16	23.367	596	35 52 16	24.413
547	29 92 09	23.388	597	35 64 09	24.434
548	30 03 04	23.409	598	35 76 04	24.454
549	30 14 01	23.431	599	35 88 01	24.474
550	30 25 00	23.452	600	36 00 00	24.495

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
601	36 12 01	24.515	651	42 38 01	25.515
602	36 24 04	24.536	652	42 51 04	25.534
603	36 36 09	24.556	653	42 64 09	25.554
604	36 48 16	24.576	654	42 77 16	25.573
605	36 60 25	24.597	655	42 90 25	25.593
606	36 72 36	24.617	656	43 03 36	25.612
607	36 84 49	24.637	657	43 16 49	25.632
608	36 96 64	24.658	658	43 29 64	25.652
609	37 08 81	24.678	659	43 42 81	25.671
610	37 21 00	24.698	660	43 56 00	25.690
611	37 33 21	24 718	661	43 69 21	25.710
612	37 45 44	24.739	662	43 82 44	25.729
613	37 57 69	24.759	663	43 95 69	25.749
614	37 69 96	24.779	664	44 08 96	25.768
615	37 82 25	24.799	665	44 22 25	25.788
616	37 94 56	24.819	666	44 35 56	25.807
617	38 06 89	24.839	667	44 48 89	25.826
618	38 19 24	24.860	668	44 62 24	25.846
619	38 31 61	24.880	669	44 75 61	25.865
620	38 44 00	24.900	670	44 89 00	25.884
621	38 56 41	24.920	671	45 02 41	25.904
622	38 68 84	24.940	672	45 15 84	25.923
623	38 81 29	24.960	673	45 29 29	25.942
624	38 93 76	24.980	674	45 42 76	25.962
625	39 06 25	25.000	675	45 56 25	25.981
626	39 18 76	25.020	676	45 69 76	26.000
627	39 31 29	25.040	677	45 83 29	26.019
628	39 43 84	25.060	678	45 96 84	26.038
629	39 56 41	25.080	679	46 10 41	26.058
630	39 69 00	25.100	680	46 24 00	26.077
631	39 81 61	25.120	681	46 37 61	26.096
632	39 94 24	25.140	682	46 51 24	26.115
633	40 06 89	25.159	683	46 64 89	26.134
634	40 19 56	25.179	684	46 78 56	26.153
635	40 32 25	25.199	685	46 92 25	26.173
636	40 44 96	25.219	686	47 05 96	26.192
637	40 57 69	25.239	687	47 19 69	26.211
638	40 70 44	25.259	688	47 33 44	26.230
639	40 83 21	25.278	689	47 47 21	26.249
640	40 96 00	25.298	690	47 61 00	26.268
641	41 08 81	25.318	691	47 74 81	26.287
642	41 21 64	25.338	692	47 88 64	26.306
643	41 34 49	25.357	693	48 02 49	26.325
644	41 47 36	25.377	694	48 16 36	26.344
645	41 60 25	25.397	695	48 30 25	26.363
646	41 73 16	25.417	696	48 44 16	26.382
647	41 86 09	25.436	697	48 58 09	26.401
648	41 99 04	25.456	698	48 72 04	26.420
649	42 12 01	25.475	699	48 86 01	26.439
650	42 25 00	25.495	700	49 00 00	26.458

Table of Squares and Square Roots—*Continued*

Number	Square	Square Root	Number	Square	Square Root
701	49 14 01	26.476	751	56 40 01	27.404
702	49 28 04	26.495	752	56 55 04	27.423
703	49 42 09	26.514	753	56 70 09	27.441
704	49 56 16	26.533	754	56 85 16	27.459
705	49 70 25	26.552	755	57 00 25	27.477
706	49 84 36	26.571	756	57 15 36	27.495
707	49 98 49	26.589	757	57 30 49	27.514
708	50 12 64	26.608	758	57 45 64	27.532
709	50 26 81	26.627	759	57 60 81	27.550
710	50 41 00	26.646	760	57 76 00	27.568
711	50 55 21	26.665	761	57 91 21	27.586
712	50 69 44	26.683	762	58 06 44	27.604
713	50 83 69	26.702	763	58 21 69	27.622
714	50 97 96	26.721	764	58 36 96	27.641
715	51 12 25	26.739	765	58 52 25	27.659
716	51 26 56	26.758	766	58 67 56	27.677
717	51 40 89	26.777	767	58 82 89	27.695
718	51 55 24	26.796	768	58 98 24	27.713
719	51 69 61	26.814	769	59 13 61	27.731
720	51 84 00	26.833	770	59 29 00	27.749
721	51 98 41	26.851	771	59 44 41	27.767
722	52 12 84	26.870	772	59 59 84	27.785
723	52 27 29	26.889	773	59 75 29	27.803
724	52 41 76	26.907	774	59 90 76	27.821
725	52 56 25	26.926	775	60 06 25	27.839
726	52 70 76	26.944	776	60 21 76	27.857
727	52 85 29	26.963	777	60 37 29	27.875
728	52 99 84	26.981	778	60 52 84	27.893
729	53 14 41	27.000	779	60 68 41	27.911
730	53 29 00	27.019	780	60 84 00	27.928
731	53 43 61	27.037	781	60 99 61	27.946
732	53 58 24	27.055	782	61 15 24	27.964
733	53 72 89	27.074	783	61 30 89	27.982
734	53 87 56	27.092	784	61 46 56	28.000
735	54 02 25	27.111	785	61 62 25	28.018
736	54 16 96	27.129	786	61 77 96	28.036
737	54 31 69	27.148	787	61 93 69	28.054
738	54 46 44	27.166	788	62 09 44	28.071
739	54 61 21	27.185	789	62 25 21	28.089
740	54 76 00	27.203	790	62 41 00	28.107
741	54 90 81	27.221	791	62 56 81	28.125
742	55 05 64	27.240	792	62 72 64	28.142
743	55 20 49	27.258	793	62 88 49	28.160
744	55 35 36	27.276	794	63 04 36	28.178
745	55 50 25	27.295	795	63 20 25	28.196
746	55 65 16	27.313	796	63 36 16	28.213
747	55 80 09	27.331	797	63 52 09	28.231
748	55 95 04	27.350	798	63 68 04	28.249
749	56 10 01	27.368	799	63 84 01	28.267
750	56 25 00	27.386	800	64 00 00	28.284

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
801	64 16 01	28.302	851	72 42 01	29.172
802	64 32 04	28.320	852	72 59 04	29.189
803	64 48 09	28.337	853	72 76 09	29.206
804	64 64 16	28.355	854	72 93 16	29.223
805	64 80 25	28.373	855	73 10 25	29.240
806	64 96 36	28.390	856	73 27 36	29.257
807	65 12 49	28.408	857	73 44 49	29.275
808	65 28 64	28.425	858	73 61 64	29.292
809	65 44 81	28.443	859	73 78 81	29.309
810	65 61 00	28.460	860	73 96 00	29.326
811	65 77 21	28.478	861	74 13 21	29.343
812	65 93 44	28.496	862	74 30 44	29.360
813	66 09 69	28.513	863	74 47 69	29.377
814	66 25 96	28.531	864	74 64 96	29.394
815	66 42 25	28.548	865	74 82 25	29.411
816	66 58 56	28.566	866	74 99 56	29.428
817	66 74 89	28.583	867	75 16 89	29.445
818	66 91 24	28.601	868	75 34 24	29.462
819	67 07 61	28.618	869	75 51 61	29.479
820	67 24 00	28.636	870	75 69 00	29.496
821	67 40 41	28.653	871	75 86 41	29.513
822	67 56 84	28.671	872	76 03 84	29.530
823	67 73 29	28.688	873	76 21 29	29.547
824	67 89 76	28.705	874	76 38 76	29.563
825	68 06 25	28.723	875	76 56 25	29.580
826	68 22 76	28.740	876	76 73 76	29.597
827	68 39 29	28.758	877	76 91 29	29.614
828	68 55 84	28.775	878	77 08 84	29.631
829	68 72 41	28.792	879	77 26 41	29.648
830	68 89 00	28.810	880	77 44 00	29.665
831	39 05 61	28.827	881	77 61 61	29.682
832	69 22 24	28.844	882	77 79 24	29.698
833	69 38 89	28.862	883	77 96 89	29.715
834	69 55 56	28.879	884	78 14 56	29.732
835	69 72 25	28.896	885	78 32 25	29.749
836	69 88 96	28.914	886	78 49 96	29.766
837	70 05 69	28.931	887	78 67 69	29.783
838	70 22 44	28.948	888	78 85 44	29.799
839	70 39 21	28.965	889	79 03 21	29.816
840	70 56 00	28.983	890	79 21 00	29.833
841	70 72 81	29.000	891	79 38 81	29.850
842	70 89 64	29.017	892	79 56 64	29.866
843	71 06 49	29.034	893	79 74 49	29.883
844	71 23 36	29.052	894	79 92 36	29.900
845	71 40 25	29.069	895	80 10 25	29.916
846	71 57 16	29.086	896	80 28 16	29.933
847	71 74 09	29.103	897	80 46 09	29.950
848	71 91 04	29.120	898	80 64 04	29.967
849	72 08 01	29.138	899	80 82 01	29.983
850	72 25 00	29.155	900	81 00 00	30.000

TABLE OF SQUARES AND SQUARE ROOTS—*Continued*

Number	Square	Square Root	Number	Square	Square Root
901	81 18 01	30.017	951	90 44 01	30.838
902	81 36 04	30.033	952	90 63 04	30.854
903	81 54 09	30.050	953	90 82 09	30.871
904	81 72 16	30.067	954	91 01 16	30.887
905	81 90 25	30.083	955	91 20 25	30.903
906	82 08 36	30.100	956	91 39 36	30.919
907	82 26 49	30.116	957	91 58 49	30.935
908	82 44 64	30.133	958	91 77 64	30.952
909	82 62 81	30.150	959	91 96 81	30.968
910	82 81 00	30.166	960	92 16 00	30.984
911	82 99 21	30.183	961	92 35 21	31.000
912	83 17 44	30.199	962	92 54 44	31.016
913	83 35 69	30.216	963	92 73 69	31.032
914	83 53 96	30.232	964	92 92 96	31.048
915	83 72 25	30.249	965	93 12 25	31.064
916	83 90 56	30.265	966	93 31 56	31.081
917	84 08 89	30.282	967	93 50 89	31.097
918	84 27 24	30.299	968	93 70 24	31.113
919	84 45 61	30.315	969	93 89 61	31.129
920	84 64 00	30.332	970	94 09 00	31.145
921	84 82 41	30.348	971	94 28 41	31.161
922	85 00 84	30.364	972	94 47 84	31.177
923	85 19 29	30.381	973	94 67 29	31.193
924	85 37 76	30.397	974	94 86 76	31.209
925	85 56 25	30.414	975	95 06 25	31.225
926	85 74 76	30.430	976	95 25 76	31.241
927	85 93 29	30.447	977	95 45 29	31.257
928	86 11 84	30.463	978	95 64 84	31.273
929	86 30 41	30.480	979	95 84 41	31.289
930	86 49 00	30.496	980	96 04 00	31.305
931	86 67 61	30.512	981	96 23 61	31.321
932	86 86 24	30.529	982	96 43 24	31.337
933	87 04 89	30.545	983	96 62 89	31.353
934	87 23 56	30.561	984	96 82 56	31.369
935	87 42 25	30.578	985	97 02 25	31.385
936	87 60 96	30.594	986	97 21 96	31.401
937	87 79 69	30.610	987	97 41 69	31.417
938	87 98 44	30.627	988	97 61 44	31.432
939	88 17 21	30.643	989	97 81 21	31.448
940	88 36 00	30.659	990	98 01 00	31.464
941	88 54 81	30.676	991	98 20 81	31.480
942	88 73 64	30.692	992	98 40 64	31.496
943	88 92 49	30.708	993	98 60 49	31.512
944	89 11 36	30.725	994	98 80 36	31.528
945	89 30 25	30.741	995	99 00 25	31.544
946	89 49 16	30.757	996	99 20 16	31.559
947	89 68 09	30.773	997	99 40 09	31.575
948	89 87 04	30.790	998	99 60 04	31.591
949	90 06 01	30.806	999	99 80 01	31.607
950	90 25 00	30.822	1000	100 00 00	31.623

INDEX